

BIONET
National Computer
Resource
for
Molecular Biology

Grant No. P41RR01685-06
Annual Progress Report
March 1, 1988 - February 28, 1989

The BIONETTM Resource is funded through a cooperative Agreement with IntelliGenetics, Inc., by the Biomedical Research Technology Program, Division of Research Resources, National Institutes of Health

BIONET Resource
c/o IntelliGenetics
700 E. El Camino Real
Mountain View, CA 94040
415-324-GENE
415-962-7300

Table of Contents

1. Title Page	1
2. Description of Program Activities	2
2.1 Scientific Subprojects	3
2.1.1 Collaborative Research and Service	3
2.1.2 Technological Research and Development	6
2.1.3 Training	8
2.2 Books, Papers, Abstracts	10
2.3 Resource Summary Table	15
3. Narrative Description	17
3.1 Summary of Research Progress	17
3.1.1 Service	18
3.1.1.1 Scientific Case Studies Using BIONET	18
3.1.1.2 Scientific Consulting: BIONET User Support	21
3.1.1.3 Service	21
3.1.1.4 PC/BIONET Communications - Distribution of the Resource	23
3.1.1.5 FASTA-MAIL program	24
3.1.1.6 New User Manual	26
3.1.1.7 Revision of the <i>Introduction to BIONET</i>	26
3.1.1.8 Additional On-line HELP EXAMPLES	27
3.1.1.9 BIONET Newsletter	27
3.1.1.10 IDEAS programs	27
3.1.2 Collaborative Research	27
3.1.2.1 Collaborative Efforts by the BIONET research staff	28
3.1.2.2 DEC-2065 Software Contributors	29
3.1.2.3 PC and Minicomputer-based Software	30
3.1.2.4 Data Contributors	32
3.1.2.5 Liaison with Other Resources	34
3.1.2.6 Bulletin Boards and Leaders	37
3.1.3 Technological Research	39
3.1.3.1 Research Efforts by the BIONET Scientist	39
3.1.3.2 FASTA-MAIL	40
3.1.3.3 Development work on the new Sun central computing facility	41
3.1.3.4 BIONET Satellite Software for VAX/VMS systems	41
3.1.3.5 The RICH program	42
3.1.3.6 BioCard - a prototype menu-driven interface for BIONET	43
3.1.3.7 XGENPUB	43
3.1.4 BIONET Training Program	44
3.1.5 Resource Facilities	46
3.1.5.1 BIONET/SUN Agreement	46
3.1.5.2 Computer Hardware and Telecommunication Networks	46
3.1.5.3 Summary Statistics on Machine Use	49
3.1.5.4 Computer Software - Core Library	61
3.1.5.5 Computer Software - System Library	61
3.1.5.6 Computer Software - Contributed Library	61
3.1.5.7 Database Library	61
3.2 Highlights	62
3.3 Administrative Changes	63
3.4 Resource Advisory Committee and Allocation of Resources	64
3.5 Dissemination of Information of Resource's Capabilities	65
3.5.1 Community Interactions and Awareness	65
3.5.2 Electronic Communications	65
3.6 IMPORTANT Suggestions and Comments	66
I. BIONET Research Publications	67
II. BIONET Training Publications	68

III. BIONET Newsletters	69
IV. BIONET Software Lending Library Catalog	70
V. BIOSCI Bulletin Board Network Information	71
VI. BIONET Training Schedules	72
VII. BIONET Computer Facilities	73
VIII. Testimonials	74

List of Figures

Figure 3-1:	Actual use of the DEC-2065 for the Month of November, 1988	51
Figure 3-2:	BIONET's Percentage of Total System Use, 12/87 - 11/88	52
Figure 3-3:	BIONET's Prime Time Use of the DEC-2065 12/87 - 11/88	54
Figure 3-4:	BIONET's Non-Prime Time Use of the DEC-2065, 12/87 - 11/88	56
Figure 3-5:	BIONET's Total Use of the DEC-2065, 12/87 - 11/88	58
Figure 3-6:	Total Telenet and Compuserve Network Use, 12/87-11/88	60

List of Tables

Table 2-1: Summary of the BIONET User Community	2
Table 3-1: Summary of Monthly Rates of Inquiries	21
Table 3-2: Summary of Monthly Rates of Questions for Programs and Databases.	22
Table 3-3: BIONET Prime Time CPU Minutes	53
Table 3-4: BIONET Prime Time Connect Hours	53
Table 3-5: BIONET Non-Prime Time CPU Minutes	55
Table 3-6: BIONET Non-Prime Time Connect Hours	55
Table 3-7: BIONET Total CPU Minutes	57
Table 3-8: BIONET Total Connect Hours	57
Table 3-9: BIONET Network Usage, Connect Hours	59

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health

Division of Research Resources
Biotechnology Resources Program
Annual Progress Report
PART I

1. PHS GRANT NUMBER:

P	4	1	R	R	0	1	6	8	5	-	0	6
---	---	---	---	---	---	---	---	---	---	---	---	---
2. TITLE OF GRANT: BIONET, National Computer Resource for
Molecular Biology
3. NAME OF RECIPIENT INSTITUTION: IntelliGenetics, Inc.
4. HEALTH PROFESSIONAL SCHOOL (If applicable): _____
5. REPORTING PERIOD:
- 5a. FROM (Month, Day, Year):

0	3	-	0	1	-	8	8
---	---	---	---	---	---	---	---
- 5b. TO (Month, Day, Year):

0	2	-	2	8	-	8	9
---	---	---	---	---	---	---	---
6. PRINCIPAL INVESTIGATOR:
- 6a. NAME: Dr. Michael J. Kelly
- 6b. TITLE: President, IntelliGenetics
- 6c. SIGNATURE: *Michael J. Kelly*
7. DATE SIGNED (Month, Day, Year): December 12, 1988
8. TELEPHONE (Include Area Code):

4	1	5	-	9	6	2	-	7	3	1	3
---	---	---	---	---	---	---	---	---	---	---	---

2. Description of Program Activities

This section of our Annual Report provides statistical information on the use of the BIONETtm Resource. The period covered is 12/87 - 11/88, to coincide with the dates of preparation of our Report and to follow our procedure of providing a full year's statistical information to compare with previous years' Reports.

Individual sections are prepared under guidelines discussed previously with BRTP staff and used in our previous Reports. We use a format for reporting the hundreds of individual Principal Investigator's use that is easy for us to generate while retaining the critical information necessary for BRTP in its internal and governmental reporting requirements. Complete research abstracts are kept at IntelliGenetics and are available upon request if needed.

The BIONET User community is divided into different classes, representing different levels of use of the computer system and staff resources, as follows:

- **Class I.** Class I users represent the Service component of the scientific community. They participate in the electronic communications facilities of BIONET (bulletin boards and electronic mail), and use the Core and Contributed Software libraries to pursue their research;
- **Class II.** Class II users represent the Collaborative component of the user community. Scientists in Class II enjoy all benefits of Class I use, and in addition contribute software and expertise to BIONET, working closely with BIONET staff. This category also includes bulletin board leaders, accounts for other related Resources (GenBank, NBRF/PIR, Dana Farber, etc.), and National Advisory Committee members.
- **Class III.** Class III accounts are for BIONET satellite communications. These accounts have the same system privileges as Class IV below but are provided free of charge.
- **Class IV.** Class IV users consist of those scientists who wish access only to the electronic communication facilities of BIONET. They are given access to the electronic mail and bulletin board facilities.
- **Class V.** Class V was implemented in response to a decision of our National Advisory Committee. Class V accounts are similar to Class IV but are for use by industrial scientists. Class V users sign an agreement stating that their communications account will be used only for scientific purposes and not for commercial advertising or other promotional purposes. Because of our ARPANET connection restrictions, users in this class do not have access to the TELNET and FTP programs which allow direct contact to other computers on the ARPANET.

Information on the number of PI's by Class is summarized in Table 2-1.

Table 2-1: Summary of the BIONET User Community

Class I	813
Class II	26
Class III	7
Class IV	19
Class V	2

Total	867

The total number of laboratories with access to BIONET, 867, is a 31 percent increase over the total of 660 presented in our last annual report! This clearly demonstrates the continuing demand for and the quality of the BIONET service. Besides having a small rate of discontinued accounts (9.6%, below) the resource has continued to grow at a remarkable rate. Between 12/87 - 11/88, 270 new labs opened accounts on BIONET, and 63 (9.6%) labs discontinued their subscriptions.

The current number of BIONET PI's (867) represents about 14% of all the NIH extramural investigators (some 6100 grants total). Actually, the total number of NIH-funded investigators may be significantly less than 6100 since many investigators hold more than one grant. Since only about half of the NIH grantees mention DNA sequence, cloning, or recombinant DNA in their grants, we would estimate that the 14% of NIH grantees that are on BIONET represents between 25 and 30% of all the scientists who could possibly make use of the BIONET facility. BIONET has considerable potential for further growth. A major factor in this continued demand is the increasing recognition by the scientific community of the need for a facility which provides rapid access to molecular biology databases and which also can serve as a hub for electronic communications. As the amount of sequence data continues to expand, the central role of BIONET will increase in value.

Considerable time and effort has been expended in developing the BIONET resource to its current state. We trust that the NIH recognizes the value of this resource to the community and is aware of the dislocation that would occur in a significant number of important research projects if any disruption occurred in the service.

2.1 Scientific Subprojects

2.1.1 Collaborative Research and Service

In the following section we report the use of the BIONET Resource for Class I-V users. The "Usage Factor" is reported as both central processor unit (cpu) time, in minutes, and connect time, in hours, for each Principal Investigator on the BIONET DEC 2065 computer. These values are the sum of all usage by the PI and his or her group members ("Sub-I's").

We note that, starting in August 1988, BIONET made available a Sun 3/280 computer to handle database searches. This was necessary because of the tremendous user demands being made on the DEC computer. Since user demand was affecting the response time on the latter system, we made access to the Sun available despite its lack of accounting software. Actual user totals may therefore be higher in many cases than listed below. We are currently working on implementing accounting software on the Sun systems and statistics on its use will be available in the near future.

We report data only on those PI groups that have used the Resource during the past 12 months. Of the 867 groups with active BIONET accounts, 693, representing about 2430 individual investigators, have accessed BIONET. Last year's access total was 530 laboratories. Thus the number of groups who have accessed the system increased by 31% as compared to last year.

If frequent users are defined as those laboratories utilizing 60 or more connect hours or 120 or more cpu minutes during the year, there were 407 groups in the "frequent user" category this year (47% of all current accounts). This represents an increase of 41% over this same category last year (288).

There are 174 accounts on the system that were inactive during the year. Most of the accounts in this category were inactive either because they were created near the end of our accounting year and not yet utilized, or because they fell into the category of complimentary (mainly foreign) accounts. Foreign users are not charged an access fee but must pay their own telecommunications charges. These accounts (now totaling 111) have been used infrequently because of the expense of international telecommunications access. However, since they utilize little of our resources, the accounts have been maintained on the system and have accumulated over the past five years.

The summary usage statistics for each laboratory group follow below. Detailed usage statistics for each individual user are maintained by the BIONET computer and are available to interested parties.

We do not report Resource staff hours nor BRTP funds allocated for individual PI's because it is impossible to allocate these rationally to such a large user community. Summary information on allocation of staff hours is given in the *Resource Summary Table*.

2.1.2 Technological Research and Development

We report on the standard form summary information for our Technological Research projects.

In past years the Resource Technology used was the DEC-2065 computer. This year most work by the staff has been performed on BIONET's new Sun computers, and accounting software for reporting CPU minutes is not currently available on those machines.

The Usage Factor was previously reported as minutes of cpu time used for the project, but is not provided this year due to the lack of usage statistics on the Sun computers.

Resource Staff Hours are based on time estimates of work reported on each project. For use in further calculations (below), hours listed for each individual in the table are divided by total annual hours per individual. This yields a fractional time or "FT" for each individual on each project.

"B RTP Funds Allocated" are calculated as the sum of the following components:

- **Actual Personnel Costs.** The personnel costs for each project are derived by multiplying the above FT for each BIONET staff person's time spent on the project times their respective annual salary plus fringe benefits; the actual personnel cost is the sum of these individual figures.
- **Consultant Costs.** The FT spent by a consultant involved in a project is multiplied times the total consulting cost for the consultant; these are summed for each project where appropriate.
- **Fraction of Awarded Funds.** The fraction of total awarded funds for each project is derived by multiplying the **fractional time** spent on each project (defined below) by the **awarded funds** (defined below). The **fractional time** is determined from the sum of hours spent on the project by the investigators listed divided by the sum of hours spent on BIONET by all investigators. For this calculation we have used the actual time spent from 12/87 through 11/88. Although this period is three months out of phase with the actual grant period, we do not think the fractional time spent will change significantly during the next three months of the current period. In computation of **awarded funds**, we include only the funds allocated in the grant categories of *Supplies*, *Travel*, and *Other Expenses*. The categories of *Personnel* and *Consultants* are accounted for in the previous two computations and the remaining grant category *Equipment* is accounted for in the *Resource Summary Table*.
- **Indirect Costs.** Indirect costs are allocated by multiplying the total awarded indirect costs for this reporting period by the fractional time devoted to each project. The fractional time is determined from the sum of hours spent on the project by all investigators divided by the sum of hours spent on BIONET by all investigators.

DDR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A									
INSTITUTION: IntelliGenetics									
GRANT NUMBER		PERIOD: March 1, 1988 to February 28, 1989							
P 4 1 R 0 1 6 8 5 - 0 6									
Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following: <div style="display: flex; justify-content: space-around;"> <input checked="" type="checkbox"/> TECHNOLOGICAL RESEARCH & DEVELOPMENT <input type="checkbox"/> COLLABORATIVE RESEARCH & SERVICE <input type="checkbox"/> TRAINING </div>									
1		2		3		4		5	
Descriptive Title (80 characters)		Sciencs Axis I	Code Axis II	a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.		Resource Technology a/	Hours Used b/	Resources Staff Hours	BRIP Funds Allocated (direct and indirect)
Abstract Comparative Sequence Analysis and Software Development. Study and classification of human Alu and KpnI sequences. Development of multiple sequence alignment editor. Protein Secondary Structure Analysis. Use of comparative sequence analysis and 3-D protein structural information to investigate determinants of protein secondary structure. RICH Program Development . Algorithm development and coding of software to search sequence databases for sequences of defined composition.		9	42, 58	a, b Jurka, Jerzy W. Ph.D. Hornq, Liang J. M.S. Maulik, Sunil Ph.D. C IntelliGenetics		Sun 3/280 Sun3/60's	N/A N/A	1246 876 7	183,603
		9	42	a, b Jurka, Jerzy W. Ph.D. Maulik, Sunil Ph.D. C IntelliGenetics		Sun 3/280 Sun3/60's	N/A N/A	560 155	64,224
		9	42	a, b Maulik, Sunil Ph.D. C IntelliGenetics		Sun 3/280 Sun3/60's	N/A N/A	556	48,685
CUMULATIVE TOTALS:									

a/ Identify Resource Technologies Used.

b/ Give Hours Resource Technologies Used.

See Instructions

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A											
INSTITUTION: IntelliGenetics, Inc.											
GRANT NUMBER		PERIOD: March 1, 1988 to February 28, 1989									
P 4 1 R R 0 1 6 8 5 - 0 6											
Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:											
<input checked="" type="checkbox"/> TECHNOLOGICAL RESEARCH & DEVELOPMENT		<input type="checkbox"/> COLLABORATIVE RESEARCH & SERVICE		<input type="checkbox"/> TRAINING							
Descriptive Title (80 characters)		Science Axis I		Science Axis II		3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.		4 USAGE FACTOR Resources Technology a/ Hours Used b/ Resource Staff Hours		5 BRIP Funds Allocated (direct and indirect)	
Abstract											
User Interface Development. Development of a prototype BIONET user interface based on Hypercard software for the Apple Macintosh computer.		9		42		a, b Maulik, Sunil Ph.D. c IntelliGenetics		N/A 318		27,845	
Network Mail Server for Sequence Database Searches. Development of FASTA-MAIL software for remote data-base searching by electronic mail.		9		42		a, b Dautricourt, J.P. Ph.D. Lear, Eliot B.S. Liebschutz, Robert B.A. Yeh, Spencer B.S. c IntelliGenetics		Sun 3/280 Sun3/60's DEC 2065 N/A N/A 233 20 30		22,627	
Central Resource Development Hardware configuration and testing for Sun Microsystems Network. Systems software development.		9		42		a, b Lear, Eliot B.S. Liebschutz, Robert B.A. Diaz, Ron B.S.		Sun 3/280 Sun3/60's N/A N/A 307 367 81		66,615	
CUMULATIVE TOTALS:											
a/ Identify Resource Technologies Used.		b/ Give Hours Resources Technologies Invest.		c/ Give Hours Resources Technologies Invest.		d/ Give Hours Resources Technologies Invest.		e/ Give Hours Resources Technologies Invest.		f/ Give Hours Resources Technologies Invest.	

DRR SCIENTIFIC SUBPROJECT FORM

PART II, SECTION A

INSTITUTION: IntelliGenetics, Inc.

PERIOD: March 1, 1988 to February 28, 1989

GRANT NUMBER P 4 1 R R 0 1 6 8 5 - 0 6

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

☒ TECHNOLOGICAL RESEARCH & DEVELOPMENT

☐ COLLABORATIVE RESEARCH & SERVICE

☐ TRAINING

Descriptive Title (80 characters) Abstract	1 Science Axis		2 Code Axis II	3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.	4 USAGE FACTOR		5 BRIP Funds Allocated (direct and indirect)
	I	II			Resource Technology a/	Hours Used b/	
Electronic Communication. Establishment of communication links between BIONET and remote computer sites. Use of ARPANET, USENET, and BITNET for networked sites. Development of communications software for mail and bulletin boards on UNIX and VAX/VMS sites. Development of international newsgroup distribution network for biologists.	9	40,42		a, b Diaz, Ron Lear, Eliot Kristofferson, David Ph.D.	Sun 3/280 Sun 3/60's MicroVAX DEC 2065	N/A 456 243 200	78,580
Data Submission software. Addition of user requested features to BIONET XGENPUB data submission program.	9	42		a, b Kanerva, Lauri	DEC 2065	N/A 45	3,679
CUMULATIVE TOTALS:	8					5767	495,858

a/ Identify Resource Technologies Used.

b/ Give Hours Resource Technologies Used.

See Instructions

2.1.3 Training

We report summary information for our Training program. The sites at which BIONET provided some level of training are named here and are discussed in more detail in the *Narrative Description* section.

The method for calculation of Usage Factors and BRTP Funds Allocated is the same as that described above under Technological Research and Development.

PART II, SECTION A

INSTITUTION: IntelliGenetics

GRANT NUMBER	P 4	1	R	R	0	1	6	8	5	-	0	6
--------------	-----	---	---	---	---	---	---	---	---	---	---	---

PERIOD: March 1, 1988 to February 28, 1989

Fill out a separate Subproject Form for Core, Collaborative or Training. Check one of the following:

TECHNOLOGICAL RESEARCH & DEVELOPMENT

X
COLLABORATIVE RESEARCH & SERVICE

TRAINING

1 Descriptive Title (80 characters)		2 Science Code Axis I Axis II		3 a. Investigator(s) Name (Last Name, First & Init.) b. Degrees c. Department d. Non-Host Inst.		4 USAGE FACTOR Resource Technology Used Resource Staff a/ b/ Hours		5 BRTF Funds Allocated (direct and indirect)
Abstract	BIONET Training Program	9	40,68	a. Johncox, Vickie B.S. Bigham, Nancy M.S. Berg, Kathryn M.P.A. Davis, Karen Ph.D. Kristofferson, David Ph.D. Maulik, Sunil Ph.D. Yeh, Spencer B.S. Benton-Vosman, Trish M.S. Swank, Beth B.S.	DEC-2065 N/A Sun 3/280 " " " " " " " " " " " "	68 3 8 99 59 181 61 9 6	41,799	
	Support of training for BIONET scientist including four in-house training sessions at IntelliGenetics, phone trainings, preparation of new training documentation, and outside demonstrations at Rutgers, the NIH, FASEB, Windsor, Ontario, and Ohio State at Wooster.			C IntelliGenetics				
CUMULATIVE TOTALS:		1					494	41,799

a/ Identify Resource Technologies Used.

b/ Give Hours Resource Technologies Used.

2.2 Books, Papers, Abstracts

We list on the next BRTP form the following research reports which detail some of BIONET's progress during this last year. Copies of these reports are available in *Appendix I*.

The second form lists papers that have trained or informed scientists about the resource. Copies of these papers are in *Appendix II*.

INSTITUTION: IntelliGenetics

REPORT PERIOD: 3/1/88 to 2/28/89

Fill out a separate form for each of the following categories: Check one.

☒ TECHNOLOGICAL RESEARCH
& DEVELOPMENT☐ COLLABORATIVE RESEARCH☐ TRAINING

Author(s)

Title of Article, Journal, Vol., Number
Pages (e.g., 44-48), Year Published.

Jurka, J. and T. Smith. (1988). Proc. Natl. Acad. Sci. USA. 85,
4775-4778. A fundamental division in the Alu family of repeated
sequences.

Faulkner, D. V. and J. Jurka. (1988). Trends Biochem. Sci. 13, 321-322.
Multiple aligned sequence editor.

Jurka, J., T. F. Smith, and D. Labuda. (1988). Nucl. Acids Res. 16,
766. Small cytoplasmic Ro RNA pseudogene and an Alu repeat in
the human α -1 globingene.

Jurka, J. and R. J. Britten. (1988). Cold Spr. Harb. Symp. abstr.
Evolution of human Alu repeats: implications for genome studies.

Holsztynska, E., D. J. Waxman, and J. Jurka. (1988). Protein
Society Mtg. abstr. Studies on rat liver cytochromes P450
using comparative sequence analysis.

Maulik, S. (1988). Protein Society Mtg. abstr. Locating amino
acid patterns in proteins by composition.

Cumulative Number Published:

Books --

Papers 3

Abstracts 3

Cumulative Number in Press:

Books --

Papers --

Abstracts --

PART II, SECTION B

GRANT NUMBER

P 4 1 R R 0 1 6 8 5 - 0 6

INSTITUTION: IntelliGenetics

REPORT PERIOD: 3/1/88 to 2/28/89

Fill out a separate form for each of the following categories: Check one.

☐ TECHNOLOGICAL RESEARCH
& DEVELOPMENT☐ COLLABORATIVE RESEARCH☒ TRAINING

Author(s)

Title of Article, Journal, Vol., Number
Pages(e.g., 44-48), Year Published.

Maulik, S. (in press). Protein Sequence and Data Analysis.
Protein Databases and Software on BIONET.

Cumulative Number Published:		Books --	Papers --	Abstracts --
Cumulative Number in Press:		Books --	Papers 1	Abstracts --

We report the publications by members of the BIONET scientific community on a version of the special form provided by BRTP. These publications have **ALL** arisen from use of BIONET, although support by BIONET and the NIH has not always been acknowledged.

The figures on *Cumulative Number Published* refer to the current year alone. We have received 44 papers that were published or are in press. The total for the three previous years was 250, bringing the overall total to 294. We note that the actual number of publications which involved the use of BIONET is undoubtedly higher because many investigators have not yet replied to our requests for reprints, and the requirement to acknowledge BIONET is not strictly followed.

INSTITUTION: IntelliGenetics
REPORT PERIOD: 3/1/88 to 2/28/89

Fill out a separate form for each of the following categories: Check one.

☐ TECHNOLOGICAL RESEARCH & DEVELOPMENT
☒ COLLABORATIVE RESEARCH
☐ TRAINING

Author(s)

Title of Article, Journal, Vol., Number
Pages(e.g., 44-48), Year Published.

Please see the following pages.

Cumulative Number Published:	42	Books --	Papers 41	Abstracts 1
Cumulative Number in Press:	2	Books ---	Papers 2	Abstracts --

References

- Akella, R., P. Arasu, and A. B. Vaidya. (1988) Molecular and Biochemical Parasitology, Vol. 30, pp. 165-174. "Molecular clones of α -tubulin genes of Plasmodium yoelii reveal an unusual feature of the carboxy terminus".
- Allison, L.A. , J. K.-C. Wong, V. D. Fitzpatrick, M. Moyle, and J. Ingels. (1988) Molecular and Cellular Biology, Vol. 8, pp. 321-329. "The C-Terminal Domain of the Largest Subunit of RNA Polymerase II of Saccharomyces cerevisiae, Drosophila melanogaster, and Mammal: a Conserved Structure with an Essential Function".
- Auger, I. E. , and C. E. Lawrence. (1988) Society of Mathematical Biology, Vol. 0092-8240, pp 1-16. " Algorithms For The Optimal Identification of Segment Neighborhoods".
- Batter, D. K. , S. R. D'Mello, L. M. Turzai, H. B. Hughes III, A. E. Gioio, and B. B. Kaplan. (1988) The Journal of Neurosciences Research, Vol. 19, pp 367-376. " The Complete Nucleotide Sequence and Structure of the Gene Encoding Bovine Phenylethanolamine N-Methyltransferase".
- Bray, S. J. , W. A. Johnson, J. Hirsh, U. Heberlein, and R. Tijian. (1988) The EMBO Journal, Vol. 7, pp. 177-188. "A cis-acting element and associated binding factor required for CNS expression of the Drosophila melanogaster dopa decarboxylase gene".
- Brayton, K. A., J. Amim, H. Qui, R. Yazdanparast, M. A. Ghatei, J. M. Polak, S. R. Bloom, and J. E. Dixon. (submitted) DNA. "Cloning, Characterization, and Sequence of a Porcine cDNA Encoding a Novel Secreted Neuronal and Endocrine Protein".
- Burns, G., T. Brown, K. Hatter, J. R. Sokatch. (1988) Eur. J. Biochem., Vol. 176, pp. 165-169. "Comparison of the amino acid sequences of the transacylase components of branched chain oxoacid dehydrogenase of Pseudomonas putida, and the pyruvate and 2-oxoglutarate dehydrogenases of Escherichia coli".
- Burns, G. , T. Brown, K. Hatter, J. M. Idriss, and J. R. Sokatch. (1988) Eur. J. Biochem., Vol. 176, pp. 311-317. "Similarity of the E1 subunits of branched-chain-oxoacid dehydrogenase from Pseudomonas putida to the corresponding subunits of mammalian branched-chain-oxoacid and pyruvate dehydrogenases".

- Burns Jr. , J. M. , T. M. Daly, A. B. Vaidya, and C. A. Long. (1988) Proc. Natl. Acad. Sci. USA, Vol. 85, pp. 602-606. "The 3' portion of the gene for a Plasmodium yoelii merozoite surface antigen encodes the epitope recognized by a protective monoclonal antibody".
- Cao, M., X. Xiao, B. Egbert, T. M. Darragh, and T. S. B. Yen. "Rapid Detection of Cutaneous Herpes Simplex Virus Infection with the Polymerase Chain Reaction". (in press 9/30/88)
- Chang, J. H., C. Tamba, S. Dumbbar, and M. O. J. Olson. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 12824-12827. "cDNA and Deduced Primary Structure of Rat Protein B23, a Nucleolar Protein Containing Highly Conserved Sequences*".
- Cohen, J. I. , R. H. Miller, B. Rosenblum, K. Denniston, J. L. Gerin, and R. H. Purcell. (1988) Virology, Vol. 162, pp. 12-20. "Sequence Comparison of Woodchuck Hepatitis Virus Replicative Forms Shows Conservation of the Geneome".
- Cooke, N. E., J. Ray, J. G. Emery, and S. A. Liehaber. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 9001-9006. "Two Distinct Species of Human Growth Hormone-variant mRNA in the Human Placenta Predict the Expression of Novel Growth Hormone Proteins".
- Dickey, L. F. , E. C. Theil, Y. H. Wang, G. E. Shulls, and I. A. Wortman III. (1988) The Journal of Biological Chemistry, Vol. 263, pp 3071-3074. "The Importance of the 3' - Untranslated Region in the Translational Control of Ferritin mRNA*"
- D'Mello , S. R. , E. P. Weisberg, M. K. Stachowiak, L. M. Turzai, A. E. Gioio, and B. B. Kaplan. (1988) The Journal of Neurosciences Research, Vol. 19, pp. 440-449. "Isolation and Nucleotide Sequence of a cDNA Clone Encoding Bovine Adrenal Tyrosine Hydroxylase: Comparative Analysis of Tyrosine Hydroxylase Gene Products".
- Jeppesen, C., B. Stebbins-Boaz, and S. A. Gerbi. (1988) Nucleic Acids Research, Vol. 16, pp. 2127-2148. "Nucleotide sequence determination and secondary structure of Xenopus U3 snRNA".
- Kokubu, F., K. Hinds, R. Litman, M. J. Shamblott, and G. W. Litman. (1988) The EMBO Journal, Vol. 7, pp. 1979-1988. "Complete structure and organization of immunoglobulin heavy chain constant region genes in a phylogenetically primitive vertebrate".
- Kokubu, F., R. Litman, M. J. Shamblott, K. Hinds, and G. W. Litman. (1988) The EMBO Journal, Vol. 7, pp. 3413-3422. "Diverse organization of immunoglobulin VH gene loci in a primitive vertebrate".

- Krawetz, S. A., W. Connor, and G. H. Dixon. (1988) DNA, Vol 6, pp. 47-57.
"Cloning of Bovine P1 Protamine cDNA and the Evolution of Vertebrate P1 Protamines".
- Krawetz, S. A., W. Connor, and G. H. Dixon. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 321-326. "Bovine Protamine Genes Contain a Single Intron".
- Krawetz, S. A., and G. A. Dixon. (1988) Journal of Molecular Evolution, Vol. 27, pp. 291-297. "Sequence Similarities of the Protamine Genes: Implications for Regulation and Evolution".
- Linskens, M. H. , and J. A. Huberman. (1988) Molecular and Cellular Biology, Vol. 8, pp 4927-4935. " Organization of Replication of Ribosomal DNA in *Saccharomyces cerevisiae*".
- Manly, K. F. , G. R. Anderson, and D. L. Stoler. (1988) Journal of Virology, Vol. 62, pp. 3540-3543. " Harvey Sarcoma Virus Genome Contains No Extensive Sequences Unrelated to Those of Other Retroviruses except *ras*".
- Mayfield, J. E., B. J. Bricker, H. Godfrey, R. M. Crosby, D. J. Knight, S. M. Halling, D. Balinsky, and L. B. Tabatabai. (1988) Gene, vol 63, pp 1-9.
"The cloning, expression, and nucleotide sequence of a gene coding for an immunogenic *Brucella abortus* protein."
- Miller, R. H. (1988) Science, Vol. 239, pp. 1420-1422. "Human Immunodeficiency Virus May Encode a Novel Protein on the Genomic DNA Plus Strand".
- Miller, R. H. (1988) Virology, Vol. 164, pp. 147-155. "Close Evolutionary Relatedness of the Hepatitis B Virus and Murine Leukemia Virus Polymerase Gene Sequence".
- Nagle, G. T., S. D. Painter, J. E. Blankenship, and A. Kurosky. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 9223-9237. "Proteolytic Processing of Egg-laying Hormone-related Precursors in *Aplysia*".
- Nagle, G. T. , S. D. Painter, J. E. Blankenship, J. V. A. Choate, and A. Kurosky. (1988) Peptides, Vol. 9, pp. 867-872. "The Bag Cell Egg-Laying Hormones of *Aplysia brasiliana* and *Aplysia californica* are Identical".
- Nees, D. W. , P. A. Stein, and R. A. Ludwig. (1988) Nucleic Acids Research, Vol. 16, pp. 9839-9853. "*The Azorhizobium caulinodans nifA* gene: dentification of upstream-activation sequences including a new element, the 'anaerobo' "

- Oka, Y., and C. A. Thomas, Jr. . (1988) Nucleic Acids Research, Vol. 15, pg. 8877-8898. "The cohering telomeres of *Oxytricha*".
- Rimsky, L., J. Hauber, M. Dukovich, M. H. Malim, A. Langlois, B. R. Cullen, and W. C. Greene. (1988) Nature, Vol. 335, pp.738-740. "Functional replacement of the HIV-1 rev protein by the HTLV-1 rex protein".
- Singh, S. V., H. Ahmad, A. Kurosky, and Y. C. Awasthi. (1988) Archives of Biochemistry and Biophysics, Vol. 264, pp. 13-22. "Purification and Characterization of Unique Glutathione S-Transferases from Human Muscle".
- Suplick, K., R. Akella, A. Saul, and A. B. Vaidya. (1988) Molecular and Biochemical Parasitology, Vol. 30, pp. 289-290. "Molecular cloning and partial sequence of a 5.8 kilobase pair repetitive DNA from *Plasmodium falciparum*".
- Sung, S. J., J. M. Bjorn Dahl, C. Y. Wang, H. T. Koa, and S. M. Fu. (1988) J. Exp. Med., Vol. 167, pp. 937-953. "Production of Tumor Necrosis Factor/Cachectin by Human T Cell Lines and Peripheral Blood T Lymphocytes Stimulated By Phorbol Myristate Acetate and Anti-CD3 Antibody".
- Upton, C., J. L. Macen, R. A. Maranchuk, A. M. DeLange, and G. McFadden. (1988) Virology, Vol. 0042-6822 , pp. 229-239. "Tumorigenic Poxviruses: Fine Analysis of the Recombination Junction in Malignant Rabbit Fibroma Virus, a Recombinant between Shope Fibroma Virus and Myxoma Virus".
- Vodkin, M. D. , and J. C. Williams. (1988) Journal of Bacteriology, Mar. 88, pp. 1227-1234. " A Heat Shock Operon in *Coxiella burnetii* Produces a Major Antigen Homologous to a Protein in Both *Mycobacteria* and *Escherichia coli*"
- Vold, B. S. , C. J. Green, N. Narasimhan, M. Strem, and J. N. Hansen. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 14485-14490. "Transcriptional Analysis of *Bacillus subtilis* rRNA-tRNA Operons".
- Wang, Y. H. , S. R. Szczekan, and E. C. Theil. (1988) Metal Ion Homeostasis: Molecular Biology and Chemistry. UCLA Symposia on Molecular and Cellular Biology. Vol. 98, D. Winge & D. Hamer, Editors, Alan R. Liss, Inc., New York.
- Weber, J. L. . (1988) Molecular and Biochemical Parasitology, Vol. 29, pp. 117-124. "Interspersed repetitive DNA from *Plasmodium falciparum*".

Weber, J. L., J. A. Lyon, R. H. Wolff, T. Hall, G. H. Lowell, and J. D. Chulay. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 11421-11425. "Primary Structure of a *Plasmodium falciparum* Malaria Antigen Located at the Merozoite Surface and within the Parasitophorous Vacuole*".

Westaway, S. K., E. M. Phizicky, and J. Ableson. (1988) The Journal of Biological Chemistry, Vol. 263, pp. 3171-3176. " Structure and Function of the Yeast tRNA Ligase Gene*".

Winkfien, R. J., R. D. Moir, S. A. Krawetz, J. Blanco, J. C. States, and G. H. Dixon. (1988) Eur. J. Biochem., Vol. 176, pp. 255-264. "A New family of repetitive, retroposon-like sequences in the genome of the rainbow trout".

Wohlrab, H., R. T. Bronson, R. C. Lu, and V. Nameth. (1988) Biomedical and Biophysical Research Communications, Vol. 154, pp. 1130-1136. "Towards a Biomarker of Mammalian Senescence: Carbonic Anhydrase III".

Xiao, X., M. Cao, T. R. Miller, Z. Y. Cao, and T. S. B. Yen. The Lancet, October 15, 1988, p. 902. Papillomavirus DNA in Cervical Carcinoma Specimens from Central China. (abstract)

2.3 Resource Summary Table

The Resource Summary Table includes the totals from the previous sections of *Technological Research and Development* and *Training*. The totals for *Collaborative Research and Service* are derived from the following components:

- **Usage Factor.** The *CPU Min. Used* is the total for BIONET CPU use found in Table III-7 since this year cpu use on the DEC was essentially entirely consumed by the user community. Technological Research projects were performed on other hardware. The total of staff hours is obtained from our accounting totals of hours spent on BIONET projects and the value for Collaborative Research and Service is likewise the remainder after subtracting time spent in other categories.
- **BRTP Funds Allocated.** This is computed as the difference between the total budget for BIONET minus the categories of Technological R+D and Training, and minus the capital equipment expenditures for the year (\$36,746) listed under Administration/Miscellaneous.

The category of *Collaborative Research and Service* includes an entry of \$169,182 in the column *Other Funds*. This is the total money invoiced over the period 12/87 - 11/88 for subscription fees (\$252,932) minus outstanding receivables of \$83,750. Each PI is asked to pay an access fee to help defray the telecommunication costs for access to BIONET; this fee is currently \$400/year. By agreement with BRTP, these access fees are not grant related income.

The balance of these fees carried forward from the previous year (as of 12/1/86) was \$156,523. After twelve additional months of collecting subscription fees (\$169,182 above) and disbursing them for telecommunication expenses (\$244,321), the balance is now \$81,384.

No facility staff computer time, work hours, or BRTP funds are allocated to the category of *Administration / Miscellaneous*; we consider such time and funds to be an integral part of the support of the other components of the Resource. We do include as Funds Allocated the \$36,746 used to purchase items of capital equipment.

The category of down time includes the sum of scheduled and unscheduled maintenance on the DEC-2065 computer. In the period 12/87 - 11/88, there was a total of 149 hours (8924 cpu minutes) of downtime:

- 2545 cpu minutes of scheduled downtime for preventive maintenance and several system-related tasks.
- 6379 minutes of downtime were due to unscheduled maintenance.

The downtime reported in the Summary Table (8032 cpu min) is 90% of the total, reflecting BIONET's allocation of 90% of the machine. Note that the **total** unscheduled maintenance of 6379 cpu minutes is only 1.2% of the total cpu time available. Considering both categories of downtime, the machine has been available for use by BIONET scientists, 98.3% of the time, 24 hours a day, seven days a week. No funds have been allocated to this category.

PART II SECTION C RESOURCE SUMMARY TABLE

GRANT NUMBER		P	4	1	R	R	0	1	6	8	5	REPORT PERIOD		3/1/88		to 2/28/89	
RESOURCE COMPONENT		Number Subproject		Number Publications		Number Investigators		USAGE FACTOR		BRP Funds Allocated \$		Resource Fees \$ Collected		Other Funds \$			
								Resource Technology	CPU/ min. used.	Resource Staff Hrs							
TECHNOLOGICAL RESEARCH & DEVELOPMENT		8		6		10		DEC 2065 Sun 3/280 Sun 3/60 MicroVax	N/A	5,767	495,858						
COLLABORATIVE RESEARCH & SERVICE		693		44		2430		DEC 2065 Sun 3/280	420,932 N/A	11,103	962,651				169,182		
TRAINING		1		1		9		DEC 2065 Sun 3/280	N/A	494	41,799						
ADMINISTRATION/ MISCELLANEOUS				see preface to this section.							36,746						
DOWN TIME									8,032								
GRAND TOTALS		702		51		2449			N/A	17,364	1,537,054				169,182		

3. Narrative Description

3.1 Summary of Research Progress

Although a very significant fraction of staff time during our fifth year was involved with planning, grant writing, and other obligations for our renewal, important progress was still made on the Resource. The following sections describe in detail our accomplishments in the several components of the BIONET Resource. Here, in brief, are some of the most notable.

- The new computer network donated by Sun Microsystems is being prepared for direct access by the user community by the BIONET systems staff (Mr. Rob Liebschutz and Mr. Eliot Lear). As of August 1988 BIONET released the FASTA-MAIL program. This provided our users on the DEC with electronic mail access to high-speed database searches on our Sun 3/280 computer. Database search times were reduced from hours to tens of minutes or less, and the average mid-day user load on the DEC 2065 dropped by a factor of about three to five since it was no longer being used for these compute-intensive tasks. FASTA-MAIL used the FASTA program obtained from Dr. William Pearson, and the mail server portion was developed by Mr. Liebschutz, Mr. Lear, and Mr. Spencer Yeh.
- Dr. Jerzy Jurka, the BIONET Scientist, has published important work in the area of repetitive DNA sequence analysis. In conjunction with this research, new functionality has been added to the Multiple Aligned Sequence Editor (MASE) by the BIONET applications programmer, Mr. Liang Jen Horng. This editor was originally developed by Dr. Jurka and Donald Faulkner at Dana Farber's Molecular Biology Computer Research Resource and then extended at BIONET over the past year. Work on the editor has also involved collaborators from the machine learning group at the University of California at Santa Cruz.
- Dr. Sunil Maulik, BIONET's Senior Scientific Consultant, has continued work on the RICH program which performs database searches for sequences of defined percent composition. Dr. Maulik has obtained some interesting preliminary results with the program. He has also been involved in developing a new Hypercard [™]-based user interface for BIONET.
- The electronic communications network was significantly enhanced. The efforts of Dr. David Kristofferson led to the formation of the international BIOSCI bulletin board network. Because scientists work on a variety of computer networks around the world, we recognized the necessity of developing a mechanism to allow all of them to communicate without the necessity of learning the peculiarities of accessing each network. We sought out computer sites on all major international networks and arranged to have parallel copies of the original BIONET bulletin boards accessible from these sites. Besides BIONET in the U.S., other major BIOSCI distribution sites are situated in England, Ireland, Sweden, and Finland. Recipients of the bulletin boards from these sites are located around the world from New Zealand and Australia, the Far East, and Israel, throughout Europe, and back to North America. The bulletin boards are available to users on the ARPANET, BITNET, EARN, Usenet, NSFnet, and JANET. Users in any particular location need only post or receive messages from their closest site. Any postings at any one site are automatically forwarded by the central BIOSCI sites to all other participants on all of the above-listed networks.
- Finally, the research conducted by BIONET's 867 laboratory groups was made significantly easier by a total revision of the BIONET documentation and the production of a new User Manual. This involved major efforts by BIONET staffer's Ms. Vickie Johncox, Mr. Spencer Yeh, and Ms. Kathryn Berg. The documentation was sent free of charge to all users on the system this past summer.

3.1.1 Service

The Service component of BIONET includes primarily Class I investigators who use the BIONET Core and Contributed program Libraries to support their research. BIONET user classes were explained above under *Description of Program Activities*. The BIONET consultants also provide support as needed by the other classes of BIONET users, but these groups are much smaller than the predominant class I group.

The computing assistance given by the staff can range from simply answering routine questions to providing sophisticated help on sequence alignments or complex database searches. Especially in the latter case the staff can make a significant contribution to the attainment of a BIONET user's research goals. Viewed in this light the distinction between "service" and "research" may be hard to discern.

Before going into the details of how the BIONET staff serves the user community we wish to provide several examples of how BIONET Class I investigators utilize the resource. These "case studies" will demonstrate the importance of the work being performed by the investigators which the BIONET staff serves.

3.1.1.1 Scientific Case Studies Using BIONET

The success of BIONET can be measured in several ways. For example, one can count the number of participating scientists, or count their publications. These numbers are interesting and impressive, but do not convey the high quality of work that is being done. It is very difficult to measure this quality objectively. We have examined the publications submitted to us, and have selected two that we feel represent some of the quality research done on the system. We present these as "case studies."

"Harvey sarcoma virus genome contains no extensive sequences unrelated to those of other retroviruses except *ras*." Kenneth F. Manly, Garth R. Anderson, and Daniel L. Stoler. *Journal of Virology* 62: 3540-3543 (1988).

Dr. Manly's laboratory has been studying the VL30 multigene elements present in rats and mice with structures similar to those of retroviruses' and retrotransposons' genomes. Transcripts of these elements are packaged as pseudotypes by type C retroviruses, and, when introduced into cells, pseudotyped VL30 RNA can lead to integration at new sites in the cell genome, suggesting that they are an entirely new class of transposable elements. Further, both the Kirsten and Harvey murine sarcoma viruses are acute transforming viruses whose genomes comprise two distinct genomic elements; a *ras* oncogene and a VL30 sequence, both of which appear to contribute directly to oncogenic activity. VL30 element transcription is strongly induced as a cellular response to anoxic stress, and studies in Dr. Manly's laboratory had previously suggested that the rat VL30 sequences incorporated into Kirsten sarcoma virus genome might directly encode the major anoxic stress protein p34, lactate dehydrogenase k.

In order to characterise the VL30 domain from Harvey sarcoma virus (HaSV) and to evaluate its similarity to other retroviral sequences, Dr. Manly and his collaborators compared translated HaSV sequences with the entire NBRF-PIR database available on BIONET using the IFIND and XFASTP database similarity searching software. Translations were done in all three reading-frames using

the PEP program. Selected sequences were evaluated for significance with the XRDF program, which compares the observed similarity score with a group of similarity scores obtained by randomizing one of the sequences many times. Additionally, HaSV RNA subsequences which correspond to the regions of peptide similarity were searched against the GenBank viral nucleic acid sequences using the XFASTN program on BIONET. Terminator codons in the HaSV RNA sequence were converted to X's to allow their acceptance by XFASTP. (The X character is treated by XFASTP as an unknown residue and is given an intermediate relatedness score).

The results of the searches yielded eight major regions of sequence similarity with (optimized) similarity scores ranging from 47 to 428 and z-scores (the alignment score for the sequence expressed as the number of standard deviations above the mean of a set of scores from the randomized sequences) ranging from 5 to 42. (Table I, pg. 3541). Three regions showed greatest similarity with *gag* sequences of feline sarcoma virus, with the remaining 5 regions showing greatest similarity with the *gag* and *pol* regions of murine leukemia viruses. Two of the regions were composites of more than one reading frame. In these cases, searches showed immediately adjacent HaSV regions in different frames matching immediately adjacent regions of viral sequences. This was interpreted to mean that a mutation in the HaSV sequence had split the original coding sequence between two or more reading frames. The sequences from the different reading frames were combined and searched against the database again, treating the two combined sequences as one.

Residues 1160 to 1420 of the VL30 showed the greatest similarity (110 residue identity out of 210) to the C-terminal region of the retroviral *pol* polyprotein, which is cleaved to yield an endonuclease. Confirmation of the amino acid sequences was found by searching for nucleic acid similarities. A 62% identity was found over a 625-base overlap with Moloney leukemia virus sequences. In addition, the nucleic acid alignments showed insertions or deletions corresponding in location to the frameshifts introduced into the peptide sequences. The similarity includes an (imperfect) copy of the C-X2-C-X4-H-X4-C motif, known to be conserved in these sequences.

Manly *et al.* conclude that the sequence comparisons suggest a distant evolutionary relationship between rat VL30 sequences and murine leukemia virus sequences. A relationship between the VL30 sequences and sequences of the retrovirus group including feline sarcoma virus and baboon endogenous virus are also suggested by the results of the database searches. Further, the likelihood that VL30 sequences directly code for an anoxic stress protein (lactate dehydrogenase k) is considered unlikely, since the searches failed to find any similarity with dehydrogenase sequences. However, their extensive similarity with endonuclease sequences suggests that they may code for a protein with that function instead.

"Structure and Function of the Yeast tRNA Ligase Gene". Shawn K. Westaway, Eric M. Phizicky, and John Abelson. *Journal of Biological Chemistry* 263: 3171-3176 (1988).

Dr. Abelson's laboratory has pioneered in the study of tRNA splicing in yeast, and in this paper they describe the DNA sequence of the entire coding region of the *Saccharomyces cerevisiae* tRNA ligase gene. tRNA ligase is one of two enzymes required for tRNA splicing in yeast. The substrates for splicing are a subset of tRNA precursors containing introns. There are no obvious conserved regions at the splice junctions (unlike the case with introns in mRNA precursors) and the only common

feature is location, which is one based removed from the 3'-end of the anticodon.

The tRNA ligase molecule itself is a single ~90-kDa polypeptide likely to contain the three separate activities required for tRNA splicing - namely phosphorylation of the 5' terminus of the 3' half-tRNA in the presence of ATP; opening of the 2',3'-cyclic phosphodiester bond of the 5' half-tRNA; and ligation of the two tRNA halves. In order to study the functional domains of this protein the gene was cloned from *S. cerevisiae* and its DNA sequence determined. Westaway *et al.* cloned into M13 phage using four unique EcoRI restriction sites to subclone plasmid pUC12-RLG. Purified phage DNA templates were sequenced by the Sanger dideoxy method by first using the M13 primer and then by priming with synthetic oligonucleotides. The 4.2 kb EcoRI fragment containing the yeast tRNA ligase gene was sequenced by obtaining a restriction map comprised of the M13 and synthetic oligonucleotide primers. Sites used were: EcoRI (site A), HpaI, ScaI, BglII, HindIII, KpnI, EcoRV, XbaI, and EcoRI (site B). M13 sequencing primer was used to obtain initial sequencing information. Oligonucleotides corresponding to unique regions of nucleotide sequence were then used as primers in directing continued synthesis along the same strand of each clone. Thus the sequence of both strands of each restriction fragment was obtained. To confirm the sequence at restriction site junctions, separate overlapping clones which spanned the junctions were sequenced.

The sequences of each fragment were entered onto BIONET using the GENED sequence editor, and the entire tRNA ligase sequence assembled using the GEL program. Prior knowledge of the initiator codon allowed immediate recognition of the tRNA ligase open reading-frame. Two further open reading-frames, ORF1 and ORF2, were also discovered using the SEQ/TRANSLATE software on BIONET. The mature tRNA ligase molecule was found to be 827 amino acids long with a molecular mass of 95.4 kDa. It is a basic protein (as expected for one involved in tRNA metabolism). Analysis using the PEP program showed that greater than 10% of the amino acids are lysines, and the net charge of the protein is +12.5 (counting histidine as +0.5). Codon usage frequencies, calculated using SEQ, show that tRNA ligase uses a large percentage of rarely used codons, consistent with the hypothesis that less abundant proteins (tRNA ligase is present in approximately 400 copies/yeast cell) do not have the bias toward preferred codons seen in highly abundant proteins. Suspicions that tRNA ligase expression might be controlled by levels of intron-containing tRNAs were proven unfounded when the codon usage of tRNA ligase relative to codons in yeast which are specifically translated by tRNAs (whose precursors contain introns) were compared using the SEQ program.

Despite the similarity in mechanism of action between yeast tRNA ligase and that of T4 RNA ligase and T4 DNA ligase, no obvious sequence similarities could be detected between the yeast tRNA ligase gene product and any of the other T4 ligases when compared with the IFIND program. Further, a screening of the entire NBRF-PIR protein sequence database using both IFIND and XFASTP revealed no significant similarities between T4 ligase and any database sequences. Other data suggesting that tRNA ligase has clearly separable domains responsible for some or all of the different activities observed during tRNA splicing should allow more subtle subsequence similarities between tRNA ligase and other proteins of similar function to be discerned.

Westaway *et al.* summarise the properties of the tRNA ligase gene product as: 1) being able to catalyze three different activities; 2) forming a splicing complex with endonuclease and tRNA precursors; and 3) being localized to a specific site at or near the inner nuclear membrane. Further studies of the gene and protein product will be necessary to characterise how these different aspects

are embodied in a single polypeptide.

3.1.1.2 Scientific Consulting: BIONET User Support

The Service component of the BIONET Resource is supported by a group of three BIONET staff members as described below. The staff interacts with the community in a variety of ways, including direct support via telephone calls, electronic mail and terminal links with individual investigators. Support is also provided through staff participation at major meetings and trade shows, at trainings, and through participation in providing on-line and printed documentation for User Manuals, program descriptions and system procedures.

We currently have a full time BIONET Scientific Consultant (Dr. Karen Davis), a full-time Applications Analyst (Mr. Spencer Yeh), and a full time Senior Scientific Consultant (Dr. Sunil Maulik). The Scientific Consultant provides direct support (telephone and e-mail assistance) to the community 75% of her time on a rotating basis. The other 25% of her time is devoted to the development of the BIONET training program and other Service projects. The Applications Analyst devotes 25% of his time to user support and the rest of his time to contributed software and database development/maintenance tasks. The Senior Scientific Consultant oversees the electronic mail responses of the Consultant and Analyst, assists them in answering more complex questions, and spends the remainder of his time participating in Technological and Collaborative Research.

3.1.1.3 Service

The Service component of the BIONET Resource includes primarily Class I investigators and takes the form of answering questions by phone, by electronic mail, and by terminal links from investigators to staff. A survey of the monthly phone, mail, and terminal links for the past year shows the different uses of the BIONET Resource by the user community. The monthly inquiry rates broken down into several categories are given below.

Table 3-1: Summary of Monthly Rates of Inquiries

Category	Number of Inquiries	Percent of Total Inquiries
Programs and Databases	164	52
TOPS20 System	54	17
Administration	28	9
Electronic Mail	25	8
PC and PC Software	20	6
Telecommunications	14	4
File Transfers	14	4
	----	----
TOTALS	319	100

The yearly total is 3813 queries, or about 15 per day (based on a 260- day work year). As can be seen from Table III-1, the largest number of inquiries--52%--concern the use of BIONET's programs and databases. This Programs-and-Databases category has been subdivided into additional scientific and program categories in Table III-2.

Table 3-2: Summary of Monthly Rates of Questions for Programs and Databases.

Category	Number of Inquiries	Percent of Total Inquiries
Database Searches and Databases	85	52
DNA and Protein Sequence Analysis	27	16
Sequence and Gel Data Entry and Manipulation	19	12
Experiment Planning and Analysis	6	4
Multi-Sequence Alignment	5	3
TOPS20 System Programs	13	8
Other	9	5
	----	----
TOTALS	164	100

As shown in Table III-2, just over half of these questions about Programs and Databases concern Databases and Database Searches. This undoubtedly reflects the convenient access to fast database-searching programs and recent versions of sequence databases on BIONET. As mentioned below, the FASTA-MAIL program, introduced this year to the BIONET system, has been especially popular because of its speed and ease of use.

Table III-2 also shows that the breakdown of the other categories of Programs-and-Databases questions is as follows: 16% on Sequence Analyses; 12% on Sequence and Gel Data Entry and Manipulation; and less than 10% on each of the other categories. Inquiries about Multi-Sequence Alignment mostly concerned the use of the IntelliGenetics' GENALIGN program and William Bains' contributed XMULTAN program. Questions about TOPS20 System Programs concerned the use of the FIND and XSEARCH programs to examine database-index files, and the use of the XGENPUB program to submit sequences to the GENBANK, EMBL, and PIR databases. Inquiries about

Experiment Planning and Analysis concerned the use of the SIZER, MAP, and CLONER programs. The category OTHER covers inquiries on the use of other contributed programs, such as Michael Zuker's BIOFLD, and other miscellaneous questions.

Returning to Table III-1, the second largest category--17%--of questions listed includes those concerning the TOPS20 operating system. These are separate from the the TOPS20 system programs listed in Table III-2. These questions concerned manipulating files and directories, using the text editors, and logging in to one's BIONET account.

The third largest category (9%) comprised BIONET Administrative questions which came directly to the consultants and were rerouted to the BIONET Administrator. The BIONET administration category consisted mostly of application requests, training session information, and manual requests. In addition to the calls listed here there are many calls directly to the BIONET Administrator which are not included in Table III-1.

The fourth largest category in Table III-1 is Electronic Mail at 8%. The Telecommunications category, at 4%, includes both inquiries concerning the procedures involved in connecting to the BIONET computer and the quality of the communications between remote users and BIONET. The remaining categories are self-explanatory.

The number of inquiries this year (3813) and the rate per day (15) are only 4% higher than last year, when there were 3700 queries, or about 14 per day. We believe that this constancy of rate reflects a balance between several factors: (1) fewer inquiries per user, due to new documentation written this year, but (2) counterbalanced by an increased number of users; and (3) a slightly greater increase in the number of new accounts. This year 270 new labs opened accounts on the system versus 238 new accounts last year, an increase of 13%. However, new documentation prepared this year includes more on-line examples of the uses of the programs in the analysis of a research project and a major revision of the *Introduction to BIONET* manual, which is sent to all new users. This documentation is described below.

Not only was the total number of inquiries approximately the same this year as last year, but also the relative ranking of the various categories of inquiries on the basis of percent of inquiries was similar. The main difference is the higher proportion this year of questions on databases and database searches.

All of the above data indicate that the consultant service is an extremely important component of the BIONET Resource. There are so many features available on BIONET that the presence of a trained expert can assist users in utilizing the resource efficiently and intelligently. In addition, the large number of questions pertaining to the databases and the database searching programs point to a major use of the Resource for large scale database searching and analyses.

3.1.1.4 PC/BIONET Communications - Distribution of the Resource

It has been clear from the beginning of operation of BIONET that the majority of the user community had access to personal computers, and that they all were looking for ways to use the PC's effectively in conjunction with BIONET. We have strongly supported this method of access, to the extent of maintaining a lending (and on-line) library of software and documentation for file transfer and terminal emulation programs. We have worked closely with the community in this way because

we recognize that distribution of the Resource would be required in order for the central DEC-2065 to support an ever-increasing number of users.

If the only use of PC's was as terminals and as a source or recipient of files transferred over the network, the net burden on BIONET would probably increase, rather than decrease. The availability of PC-based software for sequence analysis has provided scientists with a means for performing many simple analyses locally. When they need access to BIONET for a more complicated analysis, they merely log on and transfer any needed files to the DEC-2065. As this "style" of computation increase in popularity, the burden on the DEC-2065 will be reduced significantly.

To facilitate use of PC's, our Consultants provide information on file formats and the use of editors on BIONET to reformat sequence files uploaded from PC's. In addition, several PC molecular biology programs are distributed via the BIONET software lending library. These programs allow users to perform many routine analyses locally. The programs are described below under *Computer Software - Contributed Library*.

The IBM PC public domain version of BIONET's on-line EMACS editor has been furnished via the lending library. "MicroEMACS" is distributed free to users and has the virtues of producing ASCII text files compatible with the BIONET software and of utilizing essentially the same command set as the mainframe editor. As new versions become available BIONET has updated its lending library diskettes and announced the availability to users via the electronic bulletin board system. Use of MicroEMACS facilitates sequence entry on user PC's and reduces dependence on the mainframe.

BIONET has also sought to standardize file transfer protocols on the system by vigorously promoting the use of the public-domain Kermit software. New versions of Kermit have been added to the lending library as they have become available. This last year saw the inclusion of a new version of Kermit for the IBM-PC which supported Tektronix 4014 emulation and automatic login macros. The former feature allowed users to obtain graphic output of plasmid maps from the IntelliGenetics (IG) CLONER program and dot-matrix plots from the IG DDMATRIX program. The latter feature (login macros) automated the process of dialing in to the BIONET Resource over Telenet or Compuserve.

3.1.1.5 FASTA-MAIL program

On August 17, a new program was introduced on BIONET that has shortened the turnaround time on database similarity searches by a factor of 20, increased the interactive response on the BIONET DEC machine by a factor of 3, while at the same time maintaining excellent sensitivity to biologically significant similarities. This was accomplished by utilizing BIONET's in-house network capabilities to set up a remote database server on the new BIONET Sun 3/280 computer using the FASTA contributed software program from David Lipman and William Pearson. The FASTA-MAIL program was based on the earlier FASTP-MAIL project which allowed BIONET users to submit remote FASTP database searches of the NBRF/PIR database on a Sun 3/280. The FASTA-MAIL program incorporated the following enhancements: a simpler user interface, the ability to do nucleic acid database searches of GenBank, the additional capability to search SWISS-PROT, the ability to set the KTUP parameter, the use of the more sensitive FASTA algorithm instead of the older FASTP algorithm, and significantly expanded on-line documentation for using the program. Since its introduction on BIONET, the FASTA-MAIL program has been extremely well-received and is used 950 times per month, one of the most popular programs on BIONET.

Extensive work was required by the BIONET staff on the FASTA-MAIL project. The FASTA program itself had to be modified to accept IntelliGenetics' file format, to have a "brief output" option to limit the alignment output to the aligned region only, and to improve the clarity and labeling of all output. The "brief output" modification was necessary since without this modification, the program frequently created output files that were larger than the BIONET mailer could handle (256 Kbytes) in a single message. The actual user interface and remote mail server software was created by the BIONET systems programmers Eliot Lear and Rob Liebschutz. This included a user interface on the DEC to submit automatically a formatted mail message to the correct address on the Sun computer, authorization software, queueing software to provide separate protein and DNA search queues on the Sun, software for running the FASTA program non-interactively, and software for mailing the search results to the correct user account on the BIONET DEC machine. Finally, two separate on-line help topics were written to provide an explanation of the FASTA algorithm and to give a step-by-step guide to using the program and accessing the output results through the MM mail program on the DEC 2065.

The FASTA-MAIL program allows both nucleic acid searches and protein searches to be run with the same program by using different scoring matrices and program switches for the two types of searches. Since the FASTA program gains some of its speed through the use of a stripped version of the databases, FASTA formatted versions of the GenBank, NBRF/PIR, and SWISS-PROT databases were created. All together, supporting these three databases for FASTA searches currently requires an additional 30 Mbytes of disk space beyond the normal requirements for these databases. We are also planning to implement a FASTA-formatted version of the EMBL nucleic acid sequence Data Library.

The user interface for the FASTA-MAIL program is designed to be extremely easy to use. The user need only respond to four simple questions to set up the search; this only takes about fifteen seconds! Since the job is run remotely, the user can log off as soon as the job is submitted. The results are automatically deposited in his or her mail file when the search is completed. Average CPU times for the searches are 1 minute for full protein databank searches and 20 minutes for full GenBank searches. Actual turnaround times are determined by the number of jobs waiting to execute in the queue. Typical turnaround times are about 5-20 minutes for full protein databank searches and 1-4 hours for full GenBank searches. Since these CPU intensive searches would otherwise have been run directly on the BIONET DEC machine, the FASTA-MAIL program has had the positive effect of dramatically increasing the responsiveness of the DEC machine. Typical mid-day user load levels on the DEC-20 were in the range of 10 - 15 before FASTA-MAIL was implemented. Now the average is 3 - 4. This means that the BIONET DEC-20 machine appears to be running about 3 to 5 times faster to the typical user.

Finally FASTA-MAIL, which although extremely fast, is still highly sensitive to biological similarities. By using a PAM250 scoring matrix for proteins, functionally similar amino acids at corresponding positions increase the score of an alignment, and by using a joining penalty in the database scan step, sequences with insertion gaps in significant regions are still kept for optimal alignment at a later stage. This last feature is the major improvement of the FASTA algorithm over the FASTP and FASTN algorithms which sometimes dropped significant matches from consideration because of the presence of small gaps. BIONET would like to thank the authors of FASTA, David Lipman of the NIH and William Pearson of the Univ. of Virginia, for allowing the BIONET

community access to their search program and for answering questions regarding the FASTA algorithm.

3.1.1.6 New User Manual

A new *User Manual* was produced this year and distributed free of charge to all current BIONET users at the end of June 1988. Unlike the old *BIONET Training Manual* which was organized by program, the *User Manual* is organized around scientific tasks and contains examples which demonstrate the integration of several programs on BIONET to accomplish the specified tasks. The following sections are included in the manual:

- Introduction
- Mechanics (Files Structure and Operating System topics)
- Sequence Entry and Editing
- Sequence Location
- Sequence and Map Display
- Sequencing Project Management
- Sequence Translation
- Sequence Composition
- Sequence Structure
- Restriction Analysis
- Cloning Simulation
- Sequence Comparison

The new manual has been favorably received by the community as demonstrated by many positive comments received from the users by the BIONET consulting staff.

3.1.1.7 Revision of the *Introduction to BIONET*

In June, a completely revised *Introduction to BIONET* manual was sent to all subscribers along with the new *User Manual*. While the *User Manual* covers the use of scientific programs on BIONET, the *Introduction to BIONET* details aspects of accessing the system and the use of other BIONET features such as electronic communications. The *Introduction to BIONET* is given to each new lab group when they sign up for BIONET. At 100 pages in length, the new "Intro" is greatly expanded from its 70 page predecessor. A sample login session is included to assist the first time BIONET user, and a trouble-shooting chapter covers many of the common problems that might be encountered while using BIONET. Other new sections include information on contributed software and databases, utility programs, network electronic mail, and electronic submission of sequences to the national databases using the XGENPUB program. A new version of the *Introduction to BIONET* will be produced when BIONET ports over to the SUN computers which run the UNIX operating system.

3.1.1.8 Additional On-line HELP EXAMPLES

As part of the consultant's on-going effort to expand and improve the on-line documentation for BIONET, several additions have been made to the HELP EXAMPLES system of on-line examples of using various programs. The most important additions have been examples of the FASTA-MAIL program, the QUEST program, and the use of the FIND and RETRIEVE commands to locate and make personal copies of entries in the databases. The FASTA-MAIL example (on-line example no. 35) guides the user through submitting a FASTA-MAIL similarity search to the DNA or protein databases, and then accessing the search results through the MM mail program. Much of the explanation involves a discussion of the FASTA-MAIL scoring and alignment output and how to interpret the results. The QUEST example (example no. 33) takes the user through using a promoter "key" (consensus pattern) in the QUEST program to search the chimpanzee beta-globin gene nucleic acid sequence for polymeraseII promoter sites. Instructions are given for using the GUIDE command to set up the search and open a file in which to collect matched sequences, loading the promoter key pattern from IntelliGenetics' KeyBank database of consensus sequences, and running the search. The FIND and RETRIEVE example (example no. 40) takes the user through a typical sequence of locating a sequence of interest in the databases and then making a personal copy of the sequence that can be freely modified by the user. Although most database retrieval tasks can now be accomplished through the new IntelliGenetics' FINDSEQ program (see below), the FIND command is still an extremely fast method of doing keyword searches, in addition to being useful for locating sequences in databases that are not yet handled by FINDSEQ such as KEYBANK and VECTORBANK. Constant revision and expansion of the on-line help system is needed as new programs and databases are added to the BIONET system.

3.1.1.9 BIONET Newsletter

The BIONET staff began producing a hardcopy newsletter called *BIONET News*. Two issues have been mailed to the user community (April and October 1988) and copies are included in *Appendix III*.

3.1.1.10 IDEAS programs

Five protein structure prediction programs from the IDEAS (Integrated Database and Extended Analysis System) suite written by Dr. Minoru Kanehisa of the National Cancer Institute and Kyoto University were released on BIONET in February. Since the IDEAS suite only runs on VAX/VMS systems, BIONET provided remote access to a networked MicroVAX owned by IntelliGenetics on which the suite is installed. To access the programs, BIONET users run a simple program which logs them into the MicroVAX. Data files may be moved to the MicroVAX and output results files can be transmitted back to the BIONET DEC-2065 for further analysis. Since its introduction, the IDEAS suite has been used an average of 23 times per month. Further details are provided below under **PC and Minicomputer-based Software**.

3.1.2 Collaborative Research

BIONET's collaborative community is made up of several components, encompassing efforts by outside scientists working in conjunction with BIONET staff. In subsequent sections we discuss each component in more detail:

- **Collaborative Efforts by the BIONET research staff.** This covers research performed by Drs. Jurka and Maulik together with outside investigators.
- **DEC-2065 Software Contributors.** This component includes those persons who have

contributed software for use by the BIONET community on the central DEC-2065 computer;

- **PC and Minicomputer-based Software.** This component includes our efforts to gather and disseminate software of special utility to the community;
- **Data Contributors.** This component includes those persons who, together with the BIONET staff, contribute data useful to the community;
- **Liaison with Other Resources.** Several accounts have been established to promote sharing of information among molecular biology computing resources;
- **Bulletin Boards and Leaders.** This component includes those persons who have agreed to maintain bulletin boards of special interest to scientists using BIONET.

3.1.2.1 Collaborative Efforts by the BIONET research staff

The BIONET research group which is headed by Dr. Jurka and includes Dr. Maulik and Mr. Liang Jen Horng, our Applications Programmer, has engaged in collaborative research on the evolution of Alu and protein sequences. Other preliminary investigations in the area of protein structure prediction have been pursued as well.

Dr. Jurka has developed collaborative research projects with Professor Roy Britten (CalTech, abstract by Jurka and Britten at Cold Spring Harb. Symp. May, 1988), and Dr. Emile Zuckerkandl (Linus Pauling Institute, abstract by Jerzy Jurka with Emile Zuckerkandl and Jerry Latter has been presented on an International FEBS Meeting, held in Sept. 1988 in Corsica). Collaborative research with both institutions is in progress and the abstracts describing this work are provided in *Appendix I*. Dr. Jurka has also been appointed to the editorial board of the *Journal of Molecular Evolution*. He has also reviewed manuscripts for *Genomics* and *Nucleic Acids Research*.

In collaboration with Aleksandar Milosavljevic and Professor David Haussler from the University of California at Santa Cruz, Dr. Jurka and Mr. Horng worked on applications of machine learning algorithms to the classification of biological sequences. In particular, they investigated a class of models of unsupervised learning, based on the principle of cognitive economy. Using prototype algorithms based on these models they were able to successfully reproduce classification of aligned human Alu sequences (Jurka and Smith, 1988). They intend to explore the applicability of these models to the classification of unaligned sequences.

As part of an on-going research project designed to aid molecular biologists in predicting the structural and functional properties of biological molecules from their sequences, BIONET has created a "higher-order" database of protein structural characteristics. Derived from the Brookhaven database (PDB) of protein structures, the DSSP database is created using the algorithm of Kabsch and Sander to create a dictionary of secondary structure of those proteins that are known to be non-homologous at the level of sequence and structure. The dictionary describes the secondary structural patterns in terms of helix, beta-bridge, beta-ladder, turn, bend, and coil and is also annotated with comments taken directly from the principal literature citations describing the structure. These include general features of the protein as well as specific annotations of residues involved in active sites or regions, forming the core, etc. The annotations are in the format: predicate-name/argument(s)/terminator to make them easily readable by a suitable computer program. BIONET has begun preliminary investigations with the Research Institute for Advanced Computer Science (RIACS) at NASA/Ames to design a Lisp-based software package that will be able

to infer structural features from this database.

3.1.2.2 DEC-2065 Software Contributors

The following are the major contributors of software for use by the community on the DEC-2065:

FASTP, FASTN, and FASTA From Bill Pearson - BIONET users have largely migrated from the older FASTP and FASTN programs this year to the FASTA-MAIL program which was described above in the **Service** section. The user interface for FASTA-MAIL runs on the DEC 2065 but the actual database searches are run remotely on BIONET's Sun 3/280 computer. The results are returned to the user on the DEC by electronic mail. The FASTA-MAIL program is used about 950 times per month or about 32 times per day!!

BIONET users are currently using FASTP on the DEC at the rate of 266 times per month. FASTN was removed from the DEC due to the advent of FASTA-MAIL and the wish to remove compute-intensive searches from the DEC but the FASTP program (which consumes far fewer resources and can be used to search user-created databases) was maintained.

MULTAN - MULTAN is a program developed by Dr. Bill Bains for aligning multiple homologous DNA sequences. While it is limited to being able to align sequences which are at least 60% homologous, it is an extremely rapid program. Multiple sequence alignment is useful for BIONET users studying evolution and for those trying to obtain a consensus from many sequences of similar function. A new version of MULTAN has been received from Dr. Bains and will be made available when BIONET users are transferred to the Sun computers.

BIOFOLD - Three years ago Dr. Michael Zuker, from the NRC laboratory in Ottawa Ontario made BIOFOLD available as a program on the BIONET computer. This program predicts RNA secondary structure and is used an average of 7 times per month.

ALIGN - Dr. Dan Davison, while a graduate student at Stony Brook at SUNY wrote and contributed two versions of his ALIGN program. The first version runs directly on the BIONET computer and can be used to align two very long DNA sequences including sequences which are not very homologous to each other and contain large gaps. The alignments are significantly better than similar heuristic alignments obtained from the SEQ SEARCH procedure in the BIONET core programs. The alignments are not as good as obtained from the SEQ ALIGN procedure but Dr. Davison's ALIGN program is significantly faster. The ALIGN program is used an average of 31 times per month.

XPROF - Dr. Rose, at Pennsylvania State University, has contributed the DEC-VAX Fortran version of his method for calculating hydropathicity profiles for proteins based on empirical observations on the extent to which amino acid residues are found to be exposed or buried. This program was released last year on the DEC and is used an average of 6 times per month.

XALIGN - As part of BIONET's effort last year to make available multiple sequence alignment software, we prepared and released XALIGN, based on Bacon and Anderson's ALIGN program. This program, as described in Bacon, D.J. and W.F. Anderson, J. Mol. Biol. 191: 153-161, 1986, finds the best local alignments among up to five protein sequences. The program was developed to run on most Fortran77-compatible computers and has been significantly modified to run on BIONET's DEC

2065 computer. XALIGN uses a large amount of CPU time and we thus asked the BIONET Community to use it during offpeak hours only, outside of 8 - 5 PST. The program has been used an average of 6 times per month since its release.

3.1.2.3 PC and Minicomputer-based Software

PC and minicomputer-based software is available to BIONET users via the BIONET lending library. This software runs on local computers instead of the DEC 2065 and, as such, helps reduce the demands on our central resource. Other minicomputer software (IDEAS, see below) is available to BIONET users on an IntelliGenetics-owned MicroVAX.

We began the lending library concept by making Kermit (a terminal emulator and file transfer protocol from Columbia University) available to BIONET users. BIONET copies Kermit onto diskettes provided by the users and returns the disks by regular mail. (Smaller lending library programs are also available for downloading from the DEC 2065.) We currently make Kermit available for Apple II, Macintosh, IBM PC and TRS-80 model computers. We have extended our lending library to include many BIONET user developed programs and a number of utility programs that are useful for file transfer on IBM PC and Macintosh computers. Software for VAX and Sun minicomputers has also been added to the lending library. A catalog of lending library software is available in *Appendix IV*. We report below the new contributed software during this past year.

MS-Kermit version 2.30

In May, version 2.30 of the public domain MS-Kermit communications package for IBM compatible computers was added to the BIONET Lending Library for free distribution to users who send in a blank diskette. This version of Kermit is significantly improved from its predecessor in that it is able to emulate a Tektronix 4010 graphics terminal. It also allows dialup procedures to be automated in a command file. The ability to emulate a graphics terminal means that BIONET users that have the Kermit package and an IBM PC or compatible can obtain the graphics output from programs such as DDMATRIX, CLONER, HPLOT, and HCOMP over their telephone connection. If they also have a graphics printing package, they can then print these graphic displays on their local dot matrix or laser printer. BIONET took advantage of the new macro capability in MS-Kermit to create login command files that allow users to dial a local Telenet or Compuserve number, and go through the complete BIONET login procedure by simply typing a single command to the Kermit program prompt. These login command files are given to users if they obtain Kermit from BIONET. Instructions for using the package with BIONET, and the original Kermit manual are also distributed to users. Since its release 74 requests for the new version of Kermit have been handled by the BIONET administrator.

New version of MacDNA

MacDNA is a program written by Robert Schleif from the Dept. of Biochemistry, Brandeis University. An updated version of MacDNA has been made available through the BIONET lending library of software and for downloading from BIONET. MacDNA is a small (65Kbyte), self-contained application that is fast and efficient at many straightforward DNA analysis tasks such as translations, character stripping, restriction mapping, consensus searching, inverting, formatting, etc. The program is not Mac-like, i.e., it uses no mouse or dialog boxes, but the current version will

run under Multifinder. The current version has enhanced formatting and editing capabilities, and does single letter translations, not included in the original version.

OligoMutantMaker

OligoMutantMaker from Kevin Beadles of the Univ. of Calif. at San Francisco has been made available to BIONET users. OligoMutantMaker simplifies the designing and screening of oligonucleotide-directed single amino acid substitution experiments by searching for nucleotide sequences which introduce a restriction endonuclease recognition sequence into the codon substitution site of the mutant. The program utilizes the redundancy of the genetic code to generate all possible nucleotide sequences for a given amino acid substitution (including nucleotide sequences in which silent mutations are introduced into the 5' and/or 3' codons immediately adjacent to the substitution site) and determines whether any restriction endonuclease recognition sequences are present. Any nucleotide sequence containing a restriction site is displayed or printed along with relevant information about the site such as its restriction enzyme(s), the random frequency of the enzyme's recognition sequence(s), the prototype of the enzyme, isoschizomers of the enzyme, and the unit cost of the enzyme from various biochemical suppliers.

OligoMutantMaker runs on the IBM PC and a version for the Macintosh is in preparation.

New Version of SEQAID II

SEQAID II is a multifunctional PC program for DNA and protein sequence analysis from Donald Roufa and Douglas Rhoads of Kansas State University. SEQAID II's functions include editing, extracting sequences from a GenBank floppy disk release, dot matrix comparison, fragment sizer, base composition, translations, protein structure and hydropathicity, restriction site search, and locating potential exons by codon bias.

One of the new features in release 3.0 is a gateway to the GenBank database (floppy disk release) and a user-friendly interface to search and extract genetic sequence data from the database. Included is a utility, GBHEADER.EXE, which permits users to modify the GenBank format file, HDR550.GBK, to accomodate future floppy disk releases. As implemented, SEQAIDII accesses floppy disk releases of GenBank either segmented on diskettes or combined in a subdirectory of the user's harddisk.

SEQAIDII version 3.0 also contains a new module that scans anonymous nucleic acid sequences for potential protein coding regions based upon codon usage frequencies. The module permits the user to construct and amplify required codon bias tables from appropriate nucleic acid sequences (cDNAs or spliced exon sequences) and to save the tables to disk.

IDEAS - As a sizeable minority of BIONET users had been requesting the availability of programs for protein structure analysis and prediction, last year BIONET investigated the availability of programs performing these types of analyses. It was decided that the IDEAS suite (Integrated Database and Sequence Analysis System) by Minoru Kanehisa of Kyoto University (and previously the National Cancer Institute) would be an appropriate package to obtain. An older version of the IDEAS package, ported to run on the BIONET DEC-20 computer, had already been made available by Dr. Kanehisa but it lacked the newer structural analysis programs needed. Since the newer

version ran only under the VMS operating system on VAX computers, in February 1988 BIONET, in co-operation with IntelliGenetics, made available an IntelliGenetics VAX/VMS computer to BIONET users (see *Technological Research*). Accordingly, BIONET set up an account on an IntelliGenetics networked microVAX computer, and ported over the IDEAS package stripped of all programs save those pertaining to structure analysis or prediction and that would ensure an added functionality to BIONET. Complete on-line help files were written describing both the programs and the access procedures, so that BIONET users could easily review the functionality of these new programs before availing themselves of them.

The programs made available were:

- **STRALI**: The STRALI program performs a secondary structure prediction by homology alignment (Kanehisa, unpublished). It searches for local regions of similarity between a query protein sequence and one whose structure has been determined and deposited in the Brookhaven Protein Databank. The Kabsch-Sander secondary structure classification (see DSSP, below) is used to predict regions of secondary structure in the query.
- **CHOFAS**: CHOFAS is the popular Chou-Fasman (Chou,P.Y. and Fasman,G.D., Biochem.(1978) 47:251-276) algorithm for secondary structure prediction complemented by the Rose algorithm (Rose,G.D., *Nature*(1978)272: 586-590) for beta-turn prediction.
- **DELPHI**: DELPHI performs protein secondary structure prediction by Robson's method (Garnier et al.,J. Mol. Biol.(1978)120:97-120). The program makes predictions for helix, (extended) beta-sheet, reverse-turn, and coil. Decision constants, provided by the user, can affect the predictions greatly.
- **ALOM**: This program attempts to identify membrane proteins based on a discriminant analysis of hydrophobic amino acids (Klein,P. et al., in prep.). The program reports whether the sequence is likely to be integral or peripheral, together with the estimated likelihood.
- **HCOMP**: Displays on a graphics terminal the hydrophobicity profiles of two aligned protein sequences. The profile is calculated by smoothing the Nozaki-Tanford hydrophobicity scale. (Nozaki,Y. and Tanford, C. J. Biol. Chem. (1971) 246: 2211-2217).
- **HPLLOT**: Plots the distribution of hydrophobic and charged amino acids using any of four possible algorithms: Nozaki-Tanford (Nozaki,Y. and Tanford, C. J. Biol. Chem. (1971) 246: 2211-2217); Hopp-Woods (Hopp,T.P. and Woods,K.R. Proc. Natl. Acad. Sci. USA (1981) 70: 3024-3828); Eisenberg (Eisenberg et al., Proc. Natl. Acad. Sci. USA (1984) 81: 140-144); or Kyte-Doolittle (Kyte,J. and Doolittle, R.F. J. Mol. Biol. (1982) 157: 105-132).

Since its introduction the IDEAS suite has been accessed an average of 23 times per month.

3.1.2.4 Data Contributors

BIONET has always provided rapid updates to all the major collections of sequence data including GenBank and EMBL nucleotide sequence collections, and the NBRF/PIR Protein Data Bank. This last year we have continued to expand the databases on-line that are related to molecular biology. This has often involved establishing contacts with database managers and providing them with the facilities on BIONET to maintain their data collections. The following summarizes our current and projected activities in this area:

Restriction Enzyme Database. We continue to provide the community with the latest additions to the Restriction Enzyme Database, through the cooperation of BIONET and Dr. Roberts at Cold

Spring Harbor. Modifications are mailed electronically to BIONET after they are incorporated into the on-line database at CSH. In addition, we provide the community with subsets of the list of enzymes that are commercially available. These subsets have been revised using the data in the July, 1988 version of the database. The lists are made available in IntelliGenetics format for use in the SEQ and PEP programs.

The catalog sources for each of the files are listed below:

amersham.lst	Amersham (8/87)
brl.lst	Bethesda Research Laboratories (6/88)
anglian.lst	Anglian Biotechnology Ltd. (1/88)
ibi.lst	International Biotechnologies Inc (7/87)
boehringer.lst	Boehringer-Mannheim (5/88)
neb.lst	New England Biolabs (7/88)
pharmacia.lst	Pharmacia P-L Biochemicals (5/88)
promega.lst	Promega Biotec (9/87)
usb.lst	United States Biochemical Corporation (4/88)

These files are extremely useful since the full Roberts' database now consists of over 900 enzymes and most users are only interested in enzymes which they can readily access. Users can also create custom databases of enzymes using the information in the commercial enzyme lists to limit their analyses to enzymes in their lab or from their favorite suppliers. Analyses run with these shorter enzyme lists will, of course, also be correspondingly faster.

SWISS-PROT. Since June, 1987 the BIONET staff has made the SWISS-PROT protein sequence database from Amos Bairoch available on BIONET. SWISS-PROT contains data obtained from the NBRF/PIR database, data translated from the EMBL DNA sequence Data Library, as well as sequences entered in-house. To maintain SWISS-PROT on BIONET requires conversion to both the IntelliGenetics format and to the FASTA format for searching by the FASTA-MAIL program. In addition, testing is done after the conversions are made to insure the integrity of the data. Maintaining the original, IntelliGenetics, and FASTA formatted versions of SWISS-PROT currently requires over 28 Mbytes of disk storage. When first released on BIONET, SWISS-PROT release 4.0 contained 1,036,010 residues in 4387 sequences. Since then the databank has approximately doubled; the current release 8.0 contains 2,224,465 residues in 7724 sequences.

Genetic Variations of *Drosophila melanogaster* - the "Red Book". BIONET has long had the full text of Lindsley and Grell's classic work "Genetic Variations of *Drosophila melanogaster*" available on line. This was made possible by help from the author, Dan Lindsley and with the permission of the original publisher, Carnegie Institution of Washington. Dan Lindsley holds an NLM grant to update this book and has also been forwarding chapters of the book to BIONET as they are finished. Because of this, the most complete collection of *Drosophila* mutants are always available on the BIONET computer. This book can be searched using one of several different text searching programs (QUEST, FIND and XSEARCH) all of which allow searches for complex boolean relationships between search terms. This also permits BIONET users to search for all genes by name, by phenotype, by affected tissues, by genetic location and by cytological location. This size of the text is still small enough to permit a complete serial search in a matter of seconds. This year BIONET received and put on-line the revisions of the chapters describing loci and mutant alleles whose names begin with the letters A through R.

SV40 Mutant List. On July 15, BIONET made available a contributed database of the SV40 large T antigen mutants compiled and maintained by Dr. J. M. Pipas of the University of Pittsburgh. The database consists of five lists of mutants divided into the following groups: early region deletion mutants including either sequence or map information; early region point mutants including either base-pair change (and corresponding amino acid substitution) or map position; and the primary sequence of truncated T antigen mutants. Each list is annotated and has a complete set of literature references which describe the mutations listed in detail. Since its introduction on BIONET the database has been distributed to five geographically distinct regions in the U.S. and in Europe using BIONET's existing ARPANET and BITNET connections, as well as being used at the rate of 16 references per month on BIONET itself.

LiMB Database. In February, BIONET made available on-line the LiMB (Listing of Molecular Biology databases) database, created by the staff at Los Alamos National Laboratories. LiMB contains information about the contents of databases related to molecular biology as well as details of how they are maintained. It was created to facilitate the process of locating and accessing databases that the research community depends on; it is also of use to those who are doing research in designing and linking these databases. Information for each database includes the purpose of the database, contact addresses including network addresses for the database staff, the history of the database, source of the database data, literature references, cross-references to other databases, details of the computer-readable form of the database, and the size of the database.

Removal of Brookhaven database from BIONET. This year access to the Brookhaven structural database had to be curtailed due to the imposition of an \$8,500 per year fee for network access by Brookhaven. Apparently a fee of this magnitude was necessary because the database effort is largely funded through the sale of tapes. The database was continued only for in-house research by the BIONET staff at the old fee of only about \$200 per release. The low usage of the database on BIONET due to the lack of graphics capabilities on the system precluded the expenditure of \$8,500. Only the DSSP program which extracted secondary structure information was available for use on the system. Given different circumstances in the future the database may be restored to BIONET. However the expenditure of such a large sum of money for this database compared to the costs of other molecular biology databases (about 10x more expensive) could be avoided if more funds were provided by the government to Brookhaven to finance its operation.

Protein Crystallography Directory. Dr. M. M. Teeter of Boston College maintains a list of the electronic mail addresses of all protein crystallographers in the U.S., Canada, Europe, and elsewhere. A copy of the latest version of the list is periodically sent electronically to BIONET over BITnet, where it is made available to any BIONET users who wish to browse it using a free-form text searching utility called FIND available on BIONET. All a user need do is attach the appropriate network suffix (e.g., .bitnet, .janet, or .earn) in order to send electronic mail to any of the crystallographers listed in the database.

3.1.2.5 Liaison with Other Resources

Several accounts have been established on BIONET to promote interaction with other, related Resources. The following is a summary of current sites with which we can exchange information.

BRTP Mailing List. Following the February 1988 meeting at the DRR BIONET established an

electronic mailing list for the Biomedical Research Technology Program. By sending a single mail message to the address brtp@BIONET-20.bio.net it is now possible for scientists at any of the B RTP-funded resources to communicate with scientists at all other resources.

Molecular Biology Computer Research Resource. The MBCRR, at the Dana-Farber Cancer Institute at Harvard, shares information through mail delivery via the GENE account and via the bulletin board system. An MBCRR bulletin board is available on the DEC 2065. Dr. Jurka at BIONET has also continued his work on the MASE editor in collaboration with Mr. Donald Faulkner at the MBCRR.

Molecular Biology Information Resource. The MBIR at Baylor formerly communicated with BIONET through Dr. Lawrence's account on the BIONET computer. After having established an Internet connection last year, the MBIR has been on our list of BIONET newsgroup recipients.

Protein Identification Resource. BIONET has provided a bulletin board expressly for the PIR which now allows members of that database staff to communicate easily with users around the world. In addition the PIR along with GenBank and EMBL now receives automatic electronic data submissions via BIONET's XGENPUB program using the common data submission form.

GenBank. GenBank continues to utilize the GenBank bulletin board on BIONET which now has worldwide distribution through the BIOSCI bulletin board network. GenBank also receives new sequence data submissions from BIONET's XGENPUB program. XGENPUB was released in 1987 and has been used to submit a total of 202 new sequences to the databases; 166 during this reporting period (12/87 - 11/88) for an average rate of 14 per month.

EMBL Databank. The EMBL continues to utilize its bulletin board on BIONET to communicate with the scientific community. It also receives electronic data submissions from BIONET's XGENPUB program as noted above.

Pittsburgh Supercomputer Center. From August 8 - August 12, Spencer Yeh, the BIONET Applications Analyst, participated in the Biomedical Supercomputer Workshop at the Pittsburgh Supercomputing Center (PSC). The Pittsburgh Supercomputer Center utilizes a Cray Y-MP supercomputer and will be the first supercomputer center to receive a Cray 3 computer when they become available in 1990. In addition, the PSC has a three-year \$2.2 million grant from the NIH's Division of Research Resources to provide the biomedical community with supercomputing resources, training, and user support. Because of this, the PSC already has both the GenBank and PIR databases on-line. The purpose of the visit to the PSC was to establish the feasibility of providing BIONET users remote access to the PSC's Cray computer for computationally-intensive tasks.

The workshop covered topics such as using the VAX frontends to the Cray, using existing software at the PSC for biomedical research, and optimization/vectorization techniques for vectorizing FORTRAN code on the Crays. Discussions with Hugh Nicholas and David Heerfield, Scientific Specialists at the PSC, addressed important issues such as the network connectivity of the PSC for remote job submission, expected speed improvement on typical BIONET applications when ported to the Cray, time allocation schemes for Cray CPU usage, suitability of the Cray for languages other than FORTRAN, especially C code, and the projected programmer time required for porting existing

applications to the Cray environment. It was learned that the PSC is changing to a UNICOS (Cray UNIX) operating system which will facilitate systems development efforts since the BIONET systems programmers are already conversant with UNIX systems.

As a result of the discussions it was concluded that it would be feasible to allow BIONET users remote access to the PSC's Cray. However, important issues still need to be addressed such as the projected speed and reliability of the NSFnet over the next 5 years, network and Cray queuing delays for remote jobs, and the overhead involved in submitting jobs to a supercomputer. While the Cray can achieve speed increases of 10 - 100 times the performance of a SUN 3/280, the turnaround time will probably be limited by the network response from California to Pittsburgh, and by the queuing system at the PSC for batch jobs. Jobs which require less than 30 min. - 1 hr. of Sun CPU time are probably too small to warrant the overhead of remote job submission to the PSC. The workshop provided an excellent introduction to using Cray computers, and the information gained will allow BIONET to evaluate its options for handling computationally intensive tasks.

Biological Matrix Workshop. BIONET has continued to support the Matrix project. Dr. Jurka, the BIONET scientist, attended the last Matrix meeting in Washington in October and BIONET has continued to maintain a newsgroup for the Matrix project.

BIOSCI bulletin board network distribution centers. BIONET initiated collaborations with five other university sites to establish the BIOSCI bulletin board network. This is described in more detail below.

Journal Editors on BIONET. Last year BIONET established accounts for the editorial board of the *Journal of Biological Chemistry*, *CABIOS*, *Cell*, and the Washington office of *Nature*. This year an account has been opened for the *Journal of Bacteriology*. These accounts serve several purposes. First, they allow easy communication between the scientific community and the on-line editorial staff. Second, they allow the editors access to sequence analysis software which they may use in reviewing manuscripts. Third, the accounts increase awareness among editors of the advantages of electronic networking, especially in regards to the problem of data submission to the nucleic acid and protein sequence databanks. Although the usual reaction on the part of journal editors has been one of reluctance to become actively involved in the data submission process (understandably so in light of their heavy workloads), their involvement in BIONET alerts them to the availability of on-line data submission software (the BIONET XGENPUB program described above) and they can then pass this information on to other scientists.

This year Dr. Herb Tabor was trained in the use of the system by Dr. Kristofferson while in Washington for the February DRR meeting. Dr. Kristofferson also trained one of Dr. Tabor's assistants at the *J. Biol. Chem.* offices in the use of electronic mail and file transfers. Joseph Palca of the Washington office of *Nature* has actively used the system during the course of the year for electronic communications. Recently the *Journal of Bacteriology* has been provided with an account on BIONET for the purpose of advance publication of the table of contents of that journal. BIONET will establish an electronic bulletin board for this purpose.

Training of the editorial staff in the use of BIONET can usually be accomplished by direct terminal links with simultaneous verbal instruction over the telephone. In the course of about 45 minutes a person can become proficient in using the electronic mail and bulletin board facility and also learn

the use of the on-line help system. This allows them to explore other features of the system at their leisure. The BIONET consultants are also available to assist the editors when called upon.

3.1.2.6 Bulletin Boards and Leaders

The following bulletin board topics are currently available on the system.

Bulletin Board Name	Description
-----	-----
AGEING	Scientific interest group
ASK-BIONET	User queries and consultant responses
BIO-CONVERSION	Scientific interest group
BIO-MATRIX	Applications of computers to biological databases
BIONEWS (BIONET-NEWS)	General announcements
BIOTECH	Biotechnology issues
CONTRIBUTED-SOFTWARE	Information on programs contributed by users
EMBL-DATABANK	Communications about EMBL databank concerns
EMPLOYMENT	Job openings
GENBANK-BB	Communications about GenBank database matters
GENE-EXPRESSION	Scientific interest group
GENOMIC-ORGANIZATION	Scientific interest group
INFO-1100	Computer interest group
INFO-AILIST	Computer interest group
INFO-AMIGA	Computer interest group
INFO-ATARI16	Computer interest group
INFO-IBM-PC	Computer interest group
INFO-KCC	Computer interest group
INFO-KERMIT	Computer interest group
INFO-LAW	Assorted legal information
INFO-MAC	Computer interest group
INFO-MODEMS	Information about modems
INFO-NEURON	Neural network computing
INFO-SUN-SPOTS	Computer interest group
INFO-TELECOM	Telecommunications
INFO-VAX	Computer interest group
MBCRR	BBoard for MBCRR announcements
METHODS-AND-REAGENTS	For reagent exchanges and announcements about lab methods
MOLECULAR-EVOLUTION	Scientific interest group
ONCOGENES	Scientific interest group
PC-COMMUNICATIONS	Information on communications software
PC-SOFTWARE	General PC software announcements
PIR	Messages to and from the PIR database staff
PLANT-MOLECULAR-BIOLOGY	Scientific interest group
PROTEIN-ANALYSIS	Scientific interest group
RESEARCH-NEWS	General interest items about science
SCIENCE-RESOURCES	Information about funding agency policy, etc.
SWISS-PROT	Messages to and from the SWISS-PROT database staff
YEAST-GENETICS	Scientific interest group

The leaders of the individual boards are:

AGEING	Sydney Shall
ASK-BIONET	David Kristofferson
BIO-CONVERSION	Eng-Leong Foo
BIO-MATRIX	Dan Davison
BIONEWS	David Kristofferson

BIOTECH	Deba Patnaik
CONTRIBUTED-SOFTWARE	BIONET Staff
EMBL-DATABANK	Graham Cameron
	& David Kristofferson
EMPLOYMENT	David Kristofferson
GENBANK-BB	Christian Burks
	& David Benton
GENE-EXPRESSION	Bill Sofer
GENOMIC-ORGANIZATION	Tom Marr
MBCRR	Susan Russo
METHODS-AND-REAGENTS	David Kristofferson
MOLECULAR-EVOLUTION	Dan Davison
ONCOGENES	David Steffen
PC-COMMUNICATIONS	David Kristofferson
PC-SOFTWARE	Doug Brutlag
PIR	David George
PLANT-MOLECULAR-BIOLOGY	Robert Jones
PROTEIN-ANALYSIS	Amos Bairoch
RESEARCH-NEWS	Sunil Maulik
SCIENCE-RESOURCES	Michele Cimbala
SWISS-PROT	Amos Bairoch
YEAST-GENETICS	John Thompson

Note that the INFO- bulletin board material is received from information sources outside of BIONET.

This year was a particularly exciting time in the development of the BIONET bulletin board system as it became the nucleus of the new international BIOSCI bulletin board system.

Because scientists work on a variety of computer networks around the world, BIONET recognized the necessity of developing a mechanism to allow all of them to communicate without the necessity of learning the peculiarities of accessing each network. We sought out computer sites on all major international networks and arranged to have parallel copies of the original BIONET bulletin boards accessible from these sites. Besides BIONET in the U.S., other major BIOSCI distribution sites are now situated at the SERC laboratory in Daresbury, England; the University College, Dublin, Ireland; the University of Uppsala in Sweden, and the University of Helsinki in Finland. Recipients of the bulletin boards from these sites are located around the world from New Zealand and Australia, the Far East, and Israel, throughout Europe, and back to North America. The bulletin boards are available to users on the ARPANET (from BIONET), BITNET (from BIONET), EARN (from Dublin, Daresbury, Helsinki, and Uppsala), Usenet (from BIONET), NSFnet (from BIONET), and JANET (from Daresbury). Users in any particular location need only post or receive messages from their closest site. Any postings at any one site are automatically forwarded by the central BIOSCI sites to all other participants on all of the above-listed networks.

A copy of the BIOSCI information sheet mailed electronically to people who request information is provided in *Appendix V*.

The following list contains the number of messages posted to each BIOSCI board from 12/87 through 11/88. All told 956 messages were posted which was an increase of 74% over the previous year. As many molecular biologists are only now discovering the use of electronic mail and bulletin boards we

expect that these high growth rates will continue into the foreseeable future. For the last five years BIONET has looked on effort expended on electronic communications as a long-term investment. This year it is clearly starting to pay off.

Bulletin Board	Messages Posted
AGEING	0 (being established)
BIO-CONVERSION	6 (started end of 11/88)
BIO-MATRIX	46
BIONEWS	187
BIOTECH	118
CONTRIBUTED-SOFTWARE	21
EMBL-DATABANK	17
EMPLOYMENT	93
GENBANK-BB	33
GENE-EXPRESSION	14
GENOMIC-ORGANIZATION	0
METHODS-AND-REAGENTS	105
MOLECULAR-EVOLUTION	44
ONCOGENES	12
PC-COMMUNICATIONS	12
PC-SOFTWARE	57
PIR	18
PLANT-MOLECULAR-BIOLOGY	5
PROTEIN-ANALYSIS	33
RESEARCH-NEWS	72
SCIENCE-RESOURCES	51
SWISS-PROT	4
YEAST-GENETICS	8

3.1.3 Technological Research

3.1.3.1 Research Efforts by the BIONET Scientist

The activities during 1988 can be divided in the following three groups: (1) Studies on interspersed repetitive elements with emphasis on Alu and L1 subfamilies; (2) Design and development of software for sequence analysis (with emphasis on sequence extraction, multiple alignment and classification); (3) Collaborative research on evolution of Alu and protein sequences (discussed above under **Collaborative Research**).

1. The biological findings on Alu classification have been published (*Proc. Natl. Acad. Sci. USA* 85, 4775-4778, 1988 & *Nucleic Acids Res.* 16, 766, 1988 see *Appendix I*). In addition to the work on Alu sequences, classes of KpnI sequences have been discovered. A manuscript on this subject will be submitted by December 1988.
2. The following activities occurred: (a) Development of a sequence editor in collaboration with Donald Faulkner of Harvard (*Trends in Biochemical Sciences* 13, 321-322, 1988). After publication of the article, several new functions have been added to MASE upon our suggestions: enhancements on COLUMN-CORRELATION, CREATE-LOCUS, SIMILARITY-DISCARD-GAPS, RENAME-LOCUS, JUMP-TO-POSITION-ABSOLUTE, modifications of PATTERN-HIGHLIGHT, MODE-ALIGNMENT, MODE-DNA, MODE-PROTEIN. In addition, we assisted in testing of these and other functions added to MASE during last year; (b) In-house work on a sensitive algorithm for sequence classification with emphasis on Alu and KpnI repeats. Previous attempts by other investigators to classify Alu repeats based on overall sequence similarities proved unsuccessful (Bains, 1986). This was quite inevitable since the diagnostic

positions are only a small portion (1-6%) of the total number of bases in Alu sequences and the distinction between them and the statistical noise could not easily be made using standard tree analysis.

3.1.3.2 FASTA-MAIL

Because of the increasing demands on the DEC 2065 BIONET needed to find alternative computing resources. Towards the end of last year, Sun Microsystems generously donated a new central computing facility to BIONET. Although some software development is necessary before we can transfer users directly to the Sun system we were able to use it remotely from the DEC for database searches through the use of our new FASTA-MAIL program. Much of the Dec's CPU resources were being spent on genetic sequence library searches, which could take up to six hours of CPU time. In an effort to ease CPU usage, we undertook to write an interface to FASTA, a sequence search program written by William Pearson that runs under UNIX. The interface we devised is known as FASTA-MAIL.

FASTA-MAIL will take submissions received via Internet mail, queue them for batch processing, and mail the results back to the submitting address. On the DEC 2065 users run a simple program and answer a few questions about the type of search that they wish to perform. This program submits these instructions together with the query sequence data to the Sun 3/280 computer at BIONET. Currently, there are two queues on our Sun that run concurrently to handle protein and nucleic acid searches, respectively. The batch processing program processes a limited grammar at the beginning of each message so that it may pass command line arguments to FASTA.

On the Sun end a message is received by the mail processing agent known as *sendmail*, which delivers the message to a program instead of a user. In this case, the program is one that checks whether the user has requested a nucleic acid search or a protein search, and places the message in the appropriate batch queue. It should be noted that as initially implemented, the batch queues used were, in fact, printer queues because SunOS's batch facilities were too weak to support multiple queues. Sequence data would enter through the print spooling system and be processed as printing output until the output filter was called. At that point, we specified our own output filter, a program known as *fastaq*. *fastaq* would then compare the sender of the database query with a list of authorized users. Once validated, the message headers would be stripped (although the program retained the return address), and a small set of parameters are read in - *ktuple* and database name. The rest of the data is then passed to FASTA, whose output is sent directly to the UNIX mail program.

Besides allowing access to BIONET users on the DEC 2065 computer, the FASTA-MAIL program ushers in a new means of performing database searches. The program can be easily modified to accept input from **any electronic mail site in the world**. Soon we expect to allow users at other sites access, but outside jobs will be placed in a lower priority queue so as not to impact significantly on our registered users. Expansion of the computer resources available at BIONET would allow us to serve a far greater number of users **very economically**.

3.1.3.3 Development work on the new Sun central computing facility

In June of 1988, the six Sun 3/60M computers and one 3/280S computer donated by Sun Microsystems arrived at BIONET. These computers are nearly operational and are providing some service to the BIONET community, particularly via FASTA-MAIL as noted above. In addition to many smaller projects (not listed), the following work on the new system has been accomplished:

- The BIONET/BIOSCI mailing lists that were administered on the DEC-20 are now being administered on the 3/280S. This will take some of the load off of the DEC-20's already loaded mailer.
- The BIONET Sun has become the major distribution point for the BIONET/BIOSCI mailing lists, passing them onto other major Internet sites.
- As mentioned previously, FASTA-MAIL was written and implemented in August so that genetic sequence searches could be done on the Sun.
- A hierarchal help system has been implemented, and the BIONET documentation is being moved into this system.
- *newuser*, a program developed by one of our staff is being implemented so that users will configure their environments the first time they log into the Suns.
- *MM*, a mail management system has been installed. An almost identical mailer was run on the DEC-20. This will eliminate relearning the electronic mail program when users migrate to the Sun system.
- *netnews*, an electronic bulletin board system, has been installed.

We shall be begin moving users to the Suns from the DEC-20 as soon as accounting software is in place.

3.1.3.4 BIONET Satellite Software for VAX/VMS systems

The BIONET Satellite software package provides the mechanisms for VAX/VMS computer systems to exchange electronic messages with a variety of computer installations. Messages are sent using telephone connections in a standard electronic mail format. Messages delivered are compliant with the address conventions of RFC822 (the standard for the format of Internet text messages). Messages can be directed to individuals or sent to a bulletin board facility. The bulletin board system implements the Standard for Interchange of USENET messages. This standard allows the host system access into the USENET news network. The software components consist of two integrated subsystems which implement the functionality as described above. The Pascal Memo Distribution Facility (PMDf) implements the MMDF protocol which provides the interface for message transmission. NEWS, written by Geoff Huston is the system which provides general conferencing in the form of a bulletin board service on the VAX/VMS host system.

The BIONET satellite system was originally designed to connect to a DEC 2060 computer running the TOPS-20 operating system. Message transmission to VAX/VMS systems was implemented using a non-standard communications protocol, CAFARD. This protocol would then interface with a mail delivery system called Pony Express. Pony Express software is a proprietary product of SRI. The acquisition of a SUN 3/280 as the primary host system for BIONET has necessitated changes in system software. The SUN machine uses the UNIX operating system. To support the CAFARD and Pony Express systems on the SUN machine would have required extensive development and maintenance time. The alternative was to look for public domain software which would provide the original functionality of the Satellite software. The selection of PMDF and NEWS were selected, in

part, to realize these goals. The MMDF protocol has been used extensively on UNIX machines and the PMDF system provides compatibility on all current VAX/VMS systems. The USENET network has been used for years as a way to conference information between end users. The NEWS system provides both connectivity and a user interface to USENET. These two subsystems are in the public domain and have source code that is distributed. This new version of the Satellite software is now undergoing local testing. Installation and documentation procedures for VAX/VMS sites are near completion. It is anticipated that site testing will begin in early January 1989.

3.1.3.5 The RICH program

It is now known that the amino acid composition of a protein plays a major role in determining its folded state (Sheridan RP *et al.* (1985) *Biopolymers* 24: 1995-2003). A fundamental pattern-matching problem in protein sequence analysis is that of finding the largest subsequences given certain density (composition) criteria only. For instance, one may wish to find the largest region in a 500 amino acid protein that contains >20% proline and >30% cysteine residues. A related problem may be finding all protein subsequences in a database that have greater than 60% hydrophobic residues.

As a result of requests such as these from BIONET users, an algorithm has been developed that will scan a sequence and heuristically discover the largest subsequence(s) that satisfy any given density criteria. Current algorithms for finding rich regions in sequences (such as RICH in the SEQ program; Brutlag, D., Clayton J., Friedland, P., & Kedes, L.H. (1982) *Nucl. Acids Res.* 10: 279), find the first such subregion satisfying the density criteria and then expand in increments of one unit (base or amino acid) until the combined density falls below the density threshold. Following this, the region is reported, and the algorithm proceeds to find the next such region. The fundamental problem with algorithms of this type is that they fail to look far enough ahead (or back) to see if sufficiently dense regions could be incorporated with the current one in order to locate the largest subsequence. Our algorithm uses a heuristic search procedure to find all seed regions satisfying the density criteria and then attempts to link all such "seeds" together. By finding all seeds in the initial pass, the algorithm circumvents the problems of the other implementations and produces the largest subregion in the second (or following) iteration or optimizing pass.

The implementation of this algorithm, termed RICH, is near completion and will be available on BIONET in 1989. The implementation has been optimized to allow RICH to scan entire databases such as the NBRF-PIR protein sequence database and locate patterns describable by their density characteristics. Uses of the RICH program might include finding hinge structures in immunoglobulin sequences (known to be rich in prolines and cysteines; Huber, R. and Bennett, W.S. (1987) *Nature* 326: 334-335) or verifying that a protein sequence satisfies the PEST hypothesis (Rogers, S., Wells, R., & Rechsteiner, M. (1986) *Science* 234: 364-368), i.e., if its half-life is related to the density of P (proline), E (glutamic acid), S (serine), and T (threonine) residues. A preliminary search of the NBRF-PIR database using RICH has shown surprising results in the type of sequences containing one or more dense PEST regions, and further studies to verify and quantify this effect are progressing. A manuscript describing RICH is in preparation, including examples and results arising from using RICH to scan databases for density-dependent patterns.

3.1.3.6 BioCard - a prototype menu-driven interface for BIONET

BIONET had previously investigated using application toolkits to develop graphics-based, menu-driven interfaces to terminal emulators running locally on users' personal computers. The aim of these investigations was to create an interface containing all the information about BIONET as easily accessible, cross-referenced help text, and to allow users to activate software or database search queries on BIONET using simple switches/buttons which would then run command-files on the central BIONET computer. With the advent of HyperCard for the Apple Macintosh, an information management system with an built-in applications generator, the creation on icons, menus, "dialog" boxes, windows, and hypertext tailored specifically to BIONET users has become straightforward. The first HyperCard application for BIONET, termed BioCard, consists of a set of graphics screens containing icons for all of the information currently on BIONET's HELP ME system, as well as buttons representing the most commonly performed tasks on BIONET. A user of BioCard could have it automatically dial the nearest Telenet or CompuServe phone number, connect to BIONET, and then run a sequence analysis program on the BIONET computer. The user can also switch between remote (BIONET) mode, and local (BioCard) mode, thus allowing users to browse through HELP information without having to quit their current application.

Currently, BioCard consists of the complete HELP ME system of BIONET in an easily searchable Hypertext format, as well as twelve buttons which, when activated, run the most commonly automated sequence analysis functions on BIONET such as database similarity searches, restriction mapping, sequencing gel assembly, etc. BioCard is also bundles with MacKermit, a public domain terminal emulation package for the Macintosh from Columbia University. Under the Macintosh's MultiFinder operating system, users can switch between local (BioCard) mode and remote (BIONET) mode in a simple and intuitive manner.

One advantage of Hypercard over other applications generators is that it is easily customizable by the user, even one with only a rudimentary familiarity with computers. Thus BIONET users will have the opportunity to select those icons, windows, and buttons that they require and use most frequently, while discarding others of no interest to them. Further, new functionality may be created by coupling or modifying the existing application icons. Finally, adventurous users may wish to design their own interfaces using BioCard as a model. Thus BioCard serves as an opportunity for BIONET to explore the many human factors engineering aspects of interface construction. Once these factors are known (through feedback from the user community over the BIONET network), BIONET plans to develop hardware-independent user interfaces using such windowing protocols as X-Windows from MIT, a public domain windowing system that is finding acceptance with such personal computer/workstation vendors as IBM, DEC, Sun, and Apple. A preliminary version of BioCard is under testing, and a first release should be available in 1989.

3.1.3.7 XGENPUB

One of the original goals of BIONET was to aid in several of the database efforts including GenBank, EMBL and the Protein Identification Resource. Initially we felt that a major contribution that BIONET could make would be to make these databases more readily available by providing software tools for database searching and analysis. However it became clear that a great number of DNA sequences were actually being determined using the IntelliGenetics GEL program on the BIONET computer. We felt that BIONET could provide a further service to both the community and to the database efforts by developing software that would allow the scientist to annotate his sequence

according to the standard GenBank format and mail the sequence and its annotation to GenBank electronically. In 1987, BIONET completed work on and released the initial version of GENPUB, currently named XGENPUB on the DEC-2065. The GENPUB program is a forms-oriented display editor that allows a person to fill in a template based on the GenBank submission form (and which can be readily changed if the GenBank form changes) giving all the requisite data about a sequence. The program automatically inserts the sequence in the appropriate place in the form by copying the sequence from a designated file on the BIONET computer. When the form is completed a single keystroke forwards the information to both the GenBank computer at Los Alamos and the EMBL computer in Heidelberg for inclusion in the next issue of the databases. At that point the entry is verified by the GenBank and EMBL staff and if they have questions about the data they can query the author by electronic mail at BIONET. To date, GENPUB has been used on BIONET to submit 202 sequences to the databases.

This year some new features were added to the XGENPUB program:

- the ability to accept any sequence data file regardless of format;
- the ability to preview the data submission form inside the XGENPUB program prior to going into the editor and;
- additional internal help documentation.

The PIR database was also added along with GenBank and EMBL as a recipient of data submissions after a common data submission form was adopted by all three databases.

Direct electronic submission mediated by GENPUB eliminates many errors in transcription. GENPUB takes sequence files directly from the software used to determine the sequence and submits it to the database. If local PC software is used to determine a sequence, the data can be sent to BIONET using an error-checking protocol such as Kermit or Modem and forwarded to GenBank via GENPUB eliminating all transcription errors. More importantly, GENPUB recruits the scientist determining a sequence to annotate it, eliminating the problems of reading and interpreting the publication and further simplifying the job of sequence collection. This concept of recruiting the aid of the scientists performing sequencing to help build the database was fundamental to the IntelliGenetics application for the GenBank contract. Since that contract was awarded to IntelliGenetics, GenBank has undertaken to make an even more sophisticated form of this program which will include error checking on the data entered into the annotations and which will be extremely portable, running on a number of microcomputers as well as mainframes. When this program is distributed to scientists who are developing new sequences, it will markedly increase the rate and quality of the data entering the GenBank database. BIONET is proud to have served the community in this way and looks forward to even closer working relationships between itself and the database efforts. We have taken an important role in not only making databases more accessible, but have also taken an active role in helping accumulate the data as well.

3.1.4 BIONET Training Program

This year we have continued holding intensive training sessions in our in-house training facility and have also given a number of demonstrations and lectures at outside sites.

We conducted two day, in-house training sessions in March, May, July, September, and November.

These five courses served a total of 45 BIONET users. The courses continue to stress the integration of the programs to solve problems. These problems include sequence entry and editing, sequencing gel management, nucleic acid and peptide sequence analysis, database structure and sequence retrieval, collecting sequences, pattern searching, and sequence similarity searches. In November we expanded the training program to include an evening session on connecting to the BIONET computer, transferring files, using a text editor, addressing electronic mail, and accessing other BIONET programs. This additional session will be included in future in-house trainings, as well. The schedules for each training are included in *Appendix VI*. The training material for these sessions has also continued to be revised by the staff during the course of the year. A new training manual for use in the courses has been produced and continually updated.

Outside seminars and demonstrations were held at Rutgers University, the NIH, the Protein Society meeting, the FASEB meeting, the Annual Meeting of the Canadian Society of Microbiology in Windsor, Ontario, and at Ohio State University at Wooster. A lecture is also planned next February in San Francisco at the annual ASCB meeting. In addition to these events, Dr. Kristofferson helped train some of the journal staff at the offices of the *Journal of Biological Chemistry* and the Washington office of *Nature* as described above under **Liasons with Other Resources**.

Rutgers. BIONET presented a poster at the bi-annual meeting at Rutgers University's workshop titled "Computers in Molecular Biology" at the Waksman Institute/Center for Advanced Biotechnology and Medicine, New Brunswick, New Jersey, April 13-15. Approximately 50 people attended. The poster described new developments at BIONET and included hands-on demonstrations of using BIONET to solve molecular biology computing tasks. In addition, BIONET participated in a panel discussion on the future of molecular biology computing. Fourteen BIONET applications were obtained as a result of the meeting.

NIH Users' Group Meeting. BIONET, in association with IntelliGenetics, organized a User's Group meeting at the National Institutes of Health, Bethesda, Maryland on August 25. The meeting featured talks describing the BIONET resource and some of the software available on it. In depth presentations were also given on sequence similarity searches and alignments, assembling and managing sequencing fragments, and recent changes in the IntelliGenetics software package. A total of 70 people attended the all-day session.

Protein Society Meeting. BIONET was invited to present a poster and demonstration at the "Computer Workshop on Protein Analysis Software" at the Second Annual meeting of the Protein Society in San Diego, August 13-17, 1988. Over 1000 attendees were expected to be present for this meeting. Two posters were presented, titled "Protein Databases and Analysis Software on BIONET" and "Locating Amino Acid Patterns by Composition" that described the current state of research and technological development in protein analysis software at BIONET. In addition during the 3 hour workshop over 30 applications were obtained and 50+ people were given hands-on experience at logging into BIONET over the Telenet/CompuServe networks and making use of the BIONET software.

FASEB Meeting. The BIONET Applications Analyst, Spencer Yeh, attended the annual meeting of the Federation of American Societies for Experimental Biology (FASEB) in May. A portable computer was used to demonstrate the BIONET system to meeting attendees. Use of the new

Tektronix graphics display in the Kermit communications program was illustrated by modifying circular plasmids within the CLONER program. In addition to generating 39 requests for new BIONET applications, the many current BIONET subscribers attending the meeting were able to obtain personal assistance with analyzing their data on BIONET. Suggestions for improving the Resource and general comments about the usefulness of the Resource to the laboratory biologist were also conveyed to the BIONET staff.

Canadian Society for Microbiology. Dr. David Kristofferson was an invited speaker to address the members during the opening symposium of the Canadian Society for Microbiology annual meeting. The meeting occurred on June 20th in Windsor, Ontario, and the opening session, attended by over 100 scientists, was entitled "Computers in Microbiology." The features of the BIONET system were described and many stimulating questions and discussions followed the presentation.

Ohio State, Wooster. The Ohio State University at Wooster conducted a summer workshop on DNA sequencing and cloning techniques attended by 27 scientists from throughout the Mid-West. This workshop included sections on computer-aided sequence analysis. Dr. Kristofferson addressed the group on June 21st. He also gave an on-line demonstration of BIONET following the lecture during which several users acquired hands-on experience and had their questions answered about the use of the Resource.

3.1.5 Resource Facilities

Previous reports have discussed the DEC-2065 and the various software and database libraries provided by the BIONET Resource. In this section we highlight significant changes and additions to the suite of hardware and software that comprise BIONET.

3.1.5.1 BIONET/SUN Agreement

Because the DEC-2065 is rapidly becoming obsolete and because equipment funds have been limited, in the summer of 1987 we submitted a proposal to Sun Microsystems to obtain a new central computing resource for BIONET. Sun agreed to donate a central Sun 3/280 file server and six Sun 3/60 client workstations. This equipment arrived in spring and became operational in June. This BIONET initiative has saved the NIH \$150,000 in equipment costs! Development efforts on the new system were described above under **Technological Research**.

3.1.5.2 Computer Hardware and Telecommunication Networks Hardware.

The current BIONET Central Resource Machine is a Digital Equipment Corporation 2065 computer with an NI20 ethernet network interface. The ethernet interface provides access to the ARPANET and other IntelliGenetics' resources.

The hardware configuration is as follows:

KL10-E Model R Processor:

- 2 MF20/MG20 Memory controllers**
- 3 MW MG20 Memory**
- MCA25 Cache Buffer Memory**

2 RH20 Massbus Channels
 NI20 Ethernet Interface

Console and Front End Processor:

PDP-11/40 CPU, 32 KW 16 bit memory
 RX02 Dual floppy disk drives
 8 DH11 Terminal interfaces 8 * 16 TTY lines each = 128 lines
 RH11 Massbus Channel
 LP20 Line printer interface

Peripherals:

3 RP07 disk drives 111MW each
 RP06 disk drive 39MW
 372 MW Total disk storage
 TU78 1600/6250-BPI tape drive
 LP26 600 LPM Line printer

Disk space (data storage)

Public structure (PS:) disk space use on the 2065 is dynamic. The following snapshot is representative of typical usage, and is taken from December 1988.

Total disk space	433,000	(pages--222 million words)
Overhead/Common	<150,000>	(Core, System and System Support Libraries)
Swapping Space	< 25,000>	
File system Overhead	< 88,000>	(Directories and index pages)

	170,000	
 BIONET Allocation	 153,000	 (90% of the available space)
BIONET Usage 12/88	<155,000>	

Unused space	< 2,000>	(Available for BIONET growth)

We conclude from these figures that BIONET is currently using 8,000 pages more than its disk space allocation.

Sun Computers

In May of 1988, BIONET took delivery of a substantial donation of computing hardware from Sun Microsystems. The BIONET staff, especially, Eliot Lear, one of the BIONET Systems Programmers, has been working to configure the hardware and software on the new Suns, so that we may gradually move users off of the Dec-20 and on to the new Sun computers.

The current Sun hardware configuration is as follows:

Sun Microsystems 3/280S Data Center Server

24 MB ECC Main Memory
 1 Xylogics 753 3.0 MB/second SMD disk controller
 1 Xylogics 472 tape controller
 1 Systech MTI 16 Channel Async Line Multiplexor
 1 450A Sun Microsystems second Ethernet controller (first

controller included on CPU board)

Peripherals:

2 Sun 626A Fujitsu M2361A Disk Drives 550 MB each
 CDC 9720-850 Disk Drive 675 MB each
 1775 MB (1.7 GB) Total disk storage
 1 Sun 675A Fujitsu 6250/1600 BPI 1/2 inch tape drive

7 Sun Microsystems 3/60M Monochrome, diskless Workstations, each
 workstation configured with 12MB main memory

The 7 diskless Sun 3/60's and the Sun 3/280 are attached to a thin ethernet dedicated to providing file service for the diskless 3/60's. The second ethernet interface on the 3/280 is connected to the IntelliGenetics backbone ethernet. All network traffic, not destined for the diskless 3/60's, is routed via this second ethernet interface, to leave the other network free for file service to the 3/60's. A Cisco Systems Gateway controls the routing of network traffic between the BIONET diskless Sun network, two other local ethernets at Intelligenetics, the Arpanet, and the soon to be established BarrNet connection. BIONET users of the Sun system have electronic mail, telnet remote login, and ftp file transfer access to the rest of the Internet. These services are also available between the Suns and the Dec-20, until the phaseout of the Dec-20 is complete.

The Sun 3/280 and the 7 Sun 3/60's are accessible via the X.25 Public Data Networks, Telenet and CompuServ (see description of X.25 networks above). This connection is implemented by connecting serial ports on our X.25 Host Pads (described above) to ports on an existing Cisco Systems Terminal Server. A terminal server is a device which allows users of serial terminal line ports to establish login connections over the ethernet to hosts on the network. Using this mechanism, we are able to distribute BIONET users across the 3/280 and 7 3/60's (plus any additional incremental resources that might be added in the future) in a way which is transparent to the users. We are working on a load balancing system, which will ensure that each time a user connects to BIONET, he/she will be connected to the most lightly loaded computer resource available on our network.

A diagram of the entire BIONET computer system and network is included in *Appendix VII*.

Public Data Network Connection.

BIONET is accessed principally over the Telenet¹ and Compuserve Public Data Networks (PDN). An X.25 PAD (packet assembler/disassembler) is located on-site for each PDN. This is known as the Host PAD, or HPAD. It provides individual terminal ports which are cross-connected to those on the DEC-2065. The Telenet trunk line operates at 9600 baud synchronously, and the PAD converts this into up to 16 asynchronous ports whose speed is typically 1200 baud. A handshaking protocol is employed to smooth over bursts of data during the multiplexing.

Connection to multiple Public Data Networks increases geographic accessibility, since areas which are served poorly by one of the PDN's are often served well by the other. Reliability is also improved by providing alternate access when service is poor or unavailable on one of the networks. Our connection to both PDN's provided incentive to each PDN to offer us lower rates.

¹The Telenet Public Data Network is operated by U.S. Sprint.

ARPANET

BIONET maintains a connection to the Arpanet which is arranged through a DARPA-funded project with IntelliCorp. In exchange for our assistance with the mechanics of the connection to ARPANET, BIONET is able to make use of this connection for communications, especially electronic mail and file transfer. Since there are mail gateways from the ARPANET to many other communications networks, this connection greatly expands BIONET's reach-- including networks such as BITNET, EARN and CSNET in addition to the DoD Internet.

The BIONET DEC-2065 is connected to a local ethernet at IntelliGenetics via its NI20 ethernet interface. A gateway connects the local ethernet to the Arpanet gateway at IntelliCorp via two 19.2KB leased lines.

NSFNET

In order to increase the reliability and speed of access to the DARPA/NSF Internet, BIONET has plans to share a T1 (1.54 Megabits/Sec) link to Barrnet (Bay Area Regional Research Network). Barrnet is the regional NSF sponsored network for the San Francisco Bay area. We expect our Barrnet connection to be operational by early January of 1989. This takes on additional importance because in recent months DARPA has taken some initial steps in the direction of scaling down the Arpanet, and there has been talk about a complete phaseout at some unspecified time in the near future.

3.1.5.3 Summary Statistics on Machine Use

Originally BIONET had access to 50% of the cpu of DEC 2065 computer and the rest was used by IntelliGenetics and its parent company IntelliCorp. With the availability of other computer resources at IntelliGenetics this situation has now changed so that BIONET's share of the computer has risen to 90%.

The cpu cycles of the DEC-2065 computer are allocated to the user community, including BIONET, by the system's class scheduler. This scheduler is given the percentage of the machine to allocate to each class of users. Any cycles not consumed by a given class ("windfall") are available to the rest of the user community. This method was chosen so that cpu cycles not consumed by one segment of the community could be used by other segments if needed, i.e., no cpu cycles are wasted if someone needs them.

Because the class scheduler also adds to the system overhead, the number of user categories has been reduced to two. This reduces system overhead and frees up computing resources. Currently the batch queue and all users except IntelliGenetics commercial time sharing customers are in class 0 with a 90% allocation of the machine. Commercial time sharing users are in the second class and have a 10% allocation of the machine.

The actual use of the machine by the BIONET community has been on average each month 88%, substantially greater than the 50% of the total cpu cycles originally allocated for BIONET. As an example, the percentage use of the machine for the month of November, 1988 is shown in Figure III-1.

The data for BIONET's percentage of system use are plotted in histogram form in Figure III-2. This figure demonstrates that BIONET has utilized well over 50% of the total cpu cycles used on the 2065, and routinely consumes over 85% of the total cpu cycles used on the system. Last year this statistic averaged about 80%.

In the following series of tables and figures, we provide further details on the actual use of the system by the BIONET community. Looking first at use of the system in prime time (8 AM - 8 PM, M-F, PST), data for cpu time and connect hours for the indicated segments of the community are given in Tables III-3 and III-4 by month, and totals. The cpu data in Table III-3 is also plotted in histogram form in Figure III-3. BIONET prime time cpu usage is up over 45% compared to the year before!

The main conclusion derived from these data is the BIONET resource is being heavily utilized. This clearly demonstrates the demand for the BIONET service and explains why we approached Sun Microsystems to obtain additional equipment.

The total number of connect hours, prime time (Table III-4), for the category BIONET Users is up over last year by 37%. This is due both to continued growth in the user community and because two batch queues were opened during the year to run jobs around the clock on the DEC. Initially these queues were used very heavily, but the advent of the FASTA-MAIL program which utilizes the Sun 3/280 computer for database searches has almost eliminated the use of the DEC batch facility for routine database searches.

The data for non-prime time (weekends and 8 PM - 8 AM M-F) are shown in Tables III-5 and III-6, and the data on cpu time are plotted in histogram form in Figure III-4. Non-prime time cpu usage by BIONET has increased by 82% over the same figure for last year! Non-prime time cpu usage remains higher than prime time because of the courtesy of users in scheduling their DEC batch jobs to run during the evening hours. In both the prime and non-prime time periods BIONET staff cpu usage has declined compared to last year as the staff has moved their work to the Sun computers. Accounting software is not yet available for these systems.

The data for total use of the Resource by BIONET are presented in Tables III-7 and III-8 and the total cpu time is summarized in Figure III-5. Overall cpu usage is up by 65% and total connect hours have increased by 36% compared to last year.

There can be no doubt when viewing these statistics that the BIONET Resource is valued and heavily utilized by the scientific community. Impressive growth in the user community (over 200 new labs this year) and the statistics on machine use indicate that BIONET may be the most heavily utilized resource funded by the NIH.

Summary data for use of our telecommunications network are presented in Table III-9 and Figure III-6 by month for the past 12 months' use of the Telenet and Compuserve networks. Total connect hours increased by 85% over last year! Note that network "connect hours" represent actual time spent using Telenet or Compuserve while "connect hours" in previous figures include batch jobs, staff usage, and system overhead. These latter factors are accounted for as time "connected" to the computer but do not represent connections via Public Data Networks. Local use of the BIONET direct-dial access lines are also not included in Table III-9 and Figure III-6.

Figure 3-1: Actual use of the DEC-2065
for the Month of November, 1988

**DEC 2065 Actual Use
November 1988**

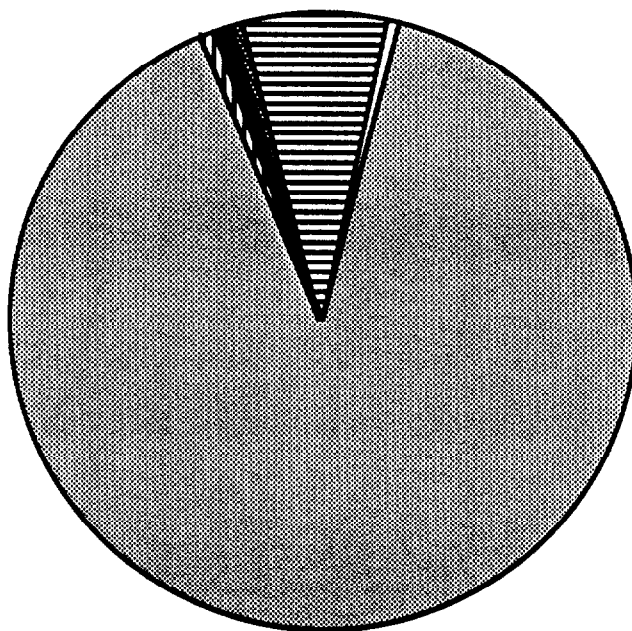
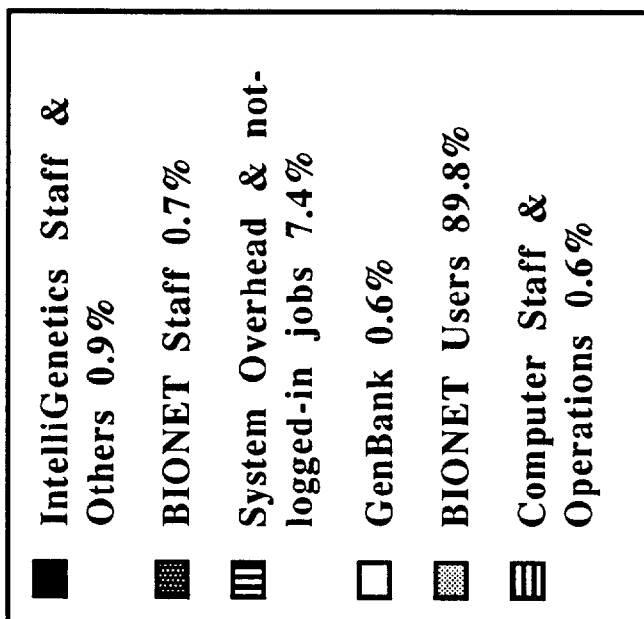
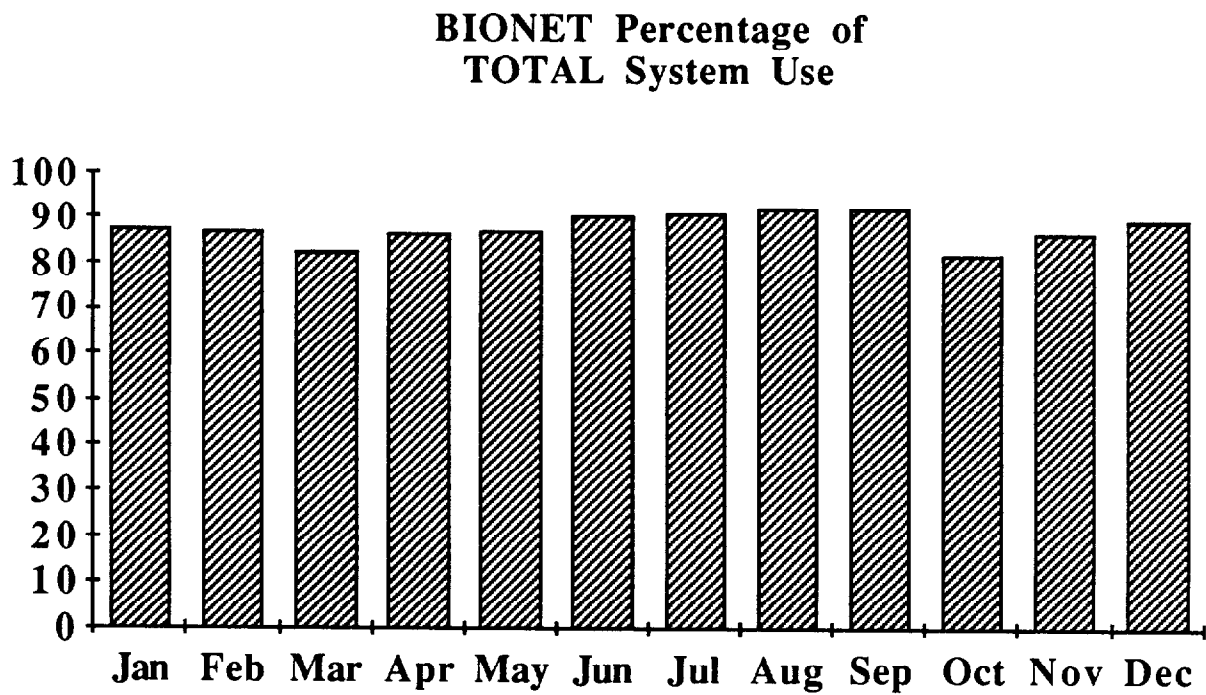


Figure 3-2: BIONET'S Percentage of Total System Use, 12/87 - 11/88



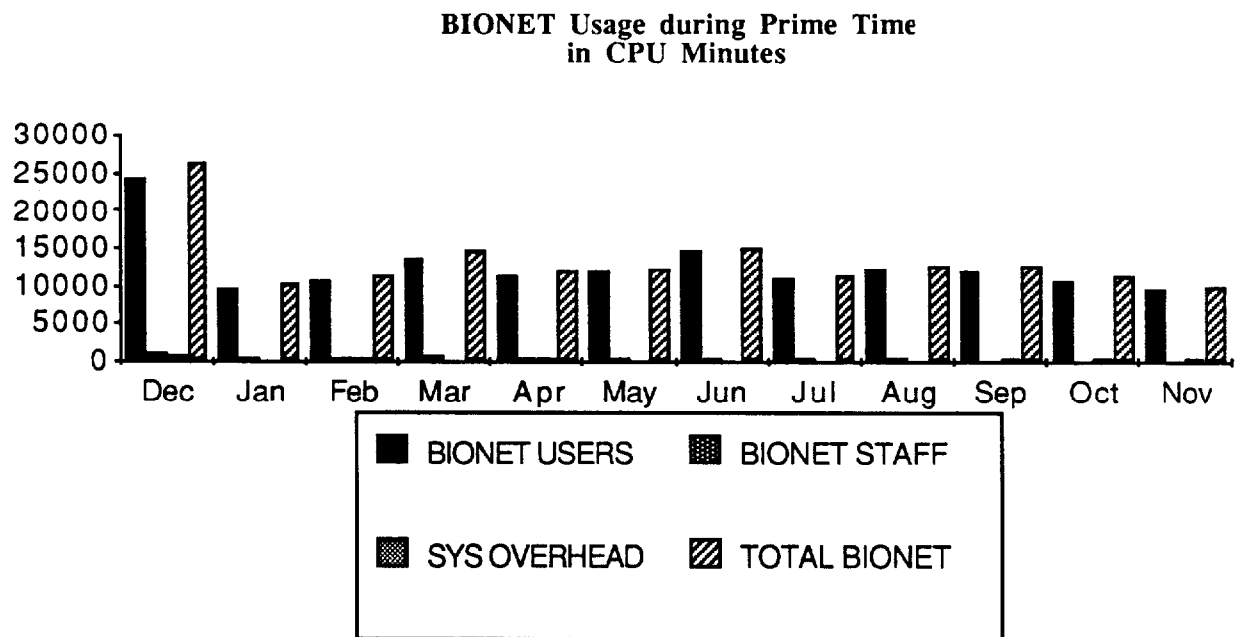
Tabel III-3: BIONET Prime Time CPU Minutes

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	24588.20	1076.90	782.60	26447.70
Jan	9840.70	414.00	169.80	10424.50
Feb	10605.80	372.90	335.25	11313.95
Mar	14010.30	649.90	251.45	14911.65
Apr	11521.40	331.10	298.70	12151.20
May	11980.10	391.50	153.45	12525.05
Jun	14747.70	338.70	229.30	15315.70
Jul	10960.00	312.50	258.00	11530.50
Aug	12371.60	288.00	167.40	12827.00
Sep	12261.40	257.90	440.05	12959.35
Oct	10897.40	231.20	322.25	11450.85
Nov	9625.80	176.90	295.45	10098.15
Total	153410.40	4841.50	3703.70	161955.60

Tabel III-4: BIONET Prime Time Connect Hours

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	7798.00	2642.90	4134.35	14575.25
Jan	2718.80	952.80	1782.45	5454.05
Feb	3658.20	867.50	1728.60	6254.30
Mar	5642.74	1257.30	2112.30	9012.34
Apr	4654.20	978.80	1701.85	7334.85
May	4391.40	957.70	1652.45	7001.55
Jun	6065.20	992.00	2040.55	9097.75
Jul	5265.30	668.20	1569.30	7502.80
Aug	5629.10	955.40	1639.50	8224.00
Sep	5168.20	866.30	1928.90	7963.40
Oct	4101.40	773.80	1629.40	6504.60
Nov	3585.20	726.50	1624.30	5936.00
Total	58677.74	12639.20	23543.95	94860.89

Figure III-3: BIONET's Prime Time Use of the DEC-2065 12/87-11/88



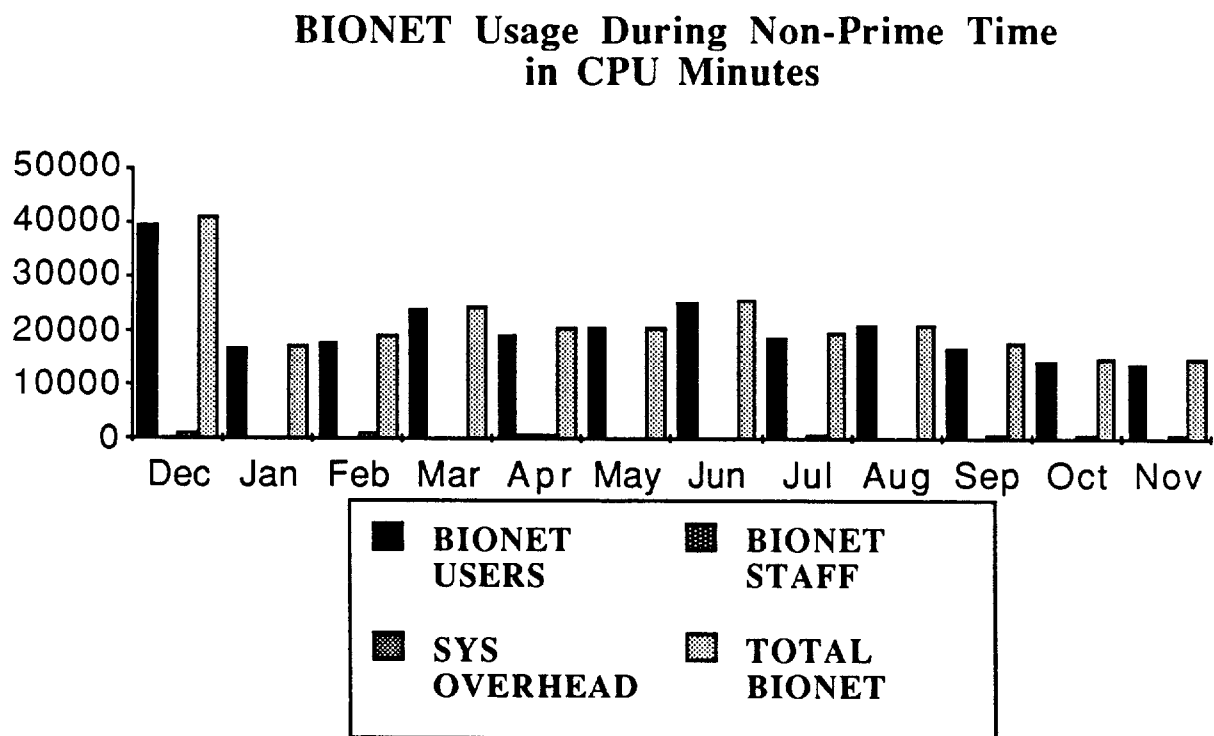
Tabel III-5: BIONET Non-Prime Time CPU Minutes

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	39571.00	322.10	1287.75	41180.85
Jan	16832.60	117.00	278.75	17228.35
Feb	17766.40	90.70	1512.10	19369.20
Mar	24050.10	482.30	235.65	24768.05
Apr	19409.10	645.70	759.55	20814.35
May	20637.40	67.90	244.65	20949.95
Jun	25492.50	77.30	267.00	25836.80
Jul	18816.00	175.60	605.25	19596.85
Aug	21007.50	20.30	177.50	21205.30
Sep	17103.50	26.70	742.45	17872.65
Oct	14466.40	16.10	809.05	15291.55
Nov	14094.70	19.20	748.70	14862.60
Total	249247.20	2060.90	7668.40	258976.50

Tabel III-6: BIONET Non-Prime Time Connect Hours

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	5439.10	450.70	5643.70	11533.50
Jan	1690.10	156.10	2640.40	4486.60
Feb	2557.80	117.60	2646.85	5322.25
Mar	4395.60	207.00	3146.75	7749.35
Apr	3170.90	199.70	2597.45	5968.05
May	3258.80	108.00	2579.00	5945.80
Jun	4045.40	87.60	3150.10	7283.10
Jul	4053.40	71.20	2248.00	6372.60
Aug	4743.40	67.70	2549.05	7360.15
Sep	3250.10	105.70	2887.10	6242.90
Oct	2553.00	33.40	2560.30	5146.70
Nov	2194.80	58.50	2524.55	4777.85
Total	41352.40	1663.20	35173.25	78188.85

Figure III-4: Bionet's Non-Prime Time Use of the Dec-2065, 12/87-11/88



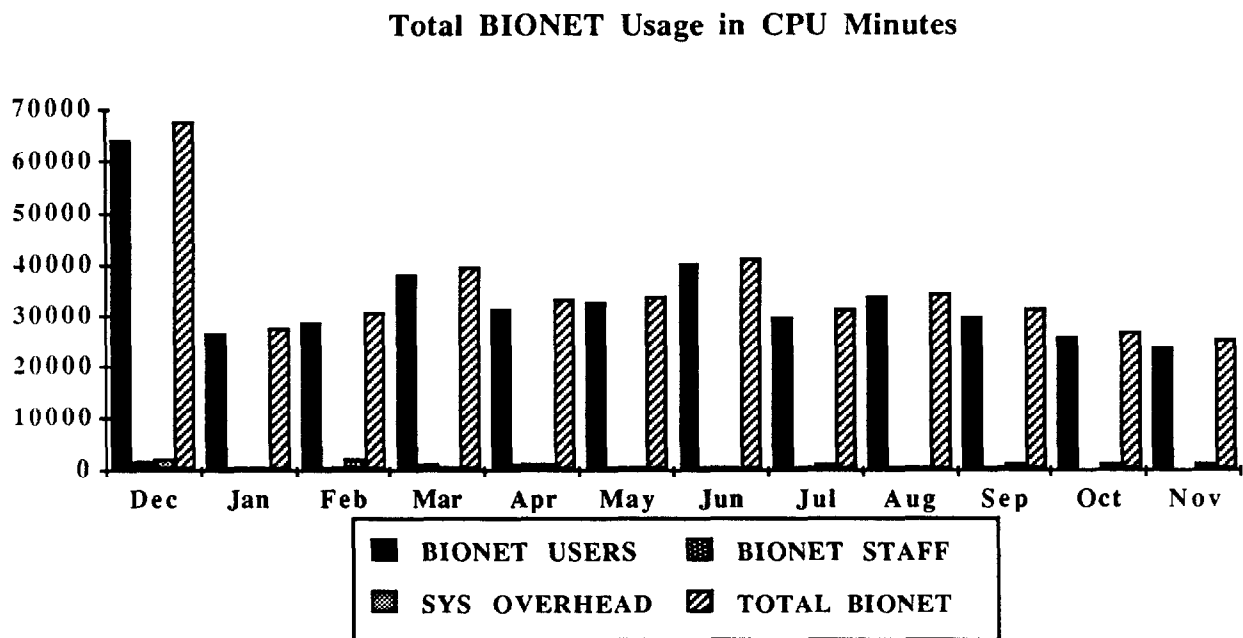
Tabel III-7: BIONET Total CPU Minutes

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	64159.20	1399.00	2070.35	67628.55
Jan	26673.30	531.00	448.55	27652.85
Feb	28372.20	463.60	1847.35	30683.15
Mar	38060.40	1132.20	487.10	39679.70
Apr	30930.50	976.80	1058.25	32965.55
May	32617.50	459.40	398.10	33475.00
Jun	40240.20	416.00	496.30	41152.50
Jul	29776.00	488.10	863.25	31127.35
Aug	33379.10	308.30	344.90	34032.30
Sep	29364.90	284.60	1182.50	30832.00
Oct	25363.80	247.30	1131.30	26742.40
Nov	23720.50	196.10	1044.15	24960.75
Total	402657.60	6902.40	11372.10	420932.10

Tabel III-8: BIONET Total Connect Hours

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	13237.10	3093.60	9778.05	26108.75
Jan	4408.90	1108.90	4422.85	9940.65
Feb	6216.00	985.10	4375.45	11576.55
Mar	10038.34	1464.30	5259.05	16761.69
Apr	7825.10	1178.50	4299.30	13302.90
May	7650.20	1065.70	4231.45	12947.35
Jun	10110.60	1079.60	5190.65	16380.85
Jul	9318.70	739.40	3817.30	13875.40
Aug	10372.50	1023.10	4188.55	15584.15
Sep	8418.30	972.00	4816.00	14206.30
Oct	6654.40	807.20	4189.70	11651.30
Nov	5780.00	785.00	4148.85	10713.85
Total	100030.14	14302.40	58717.20	173049.74

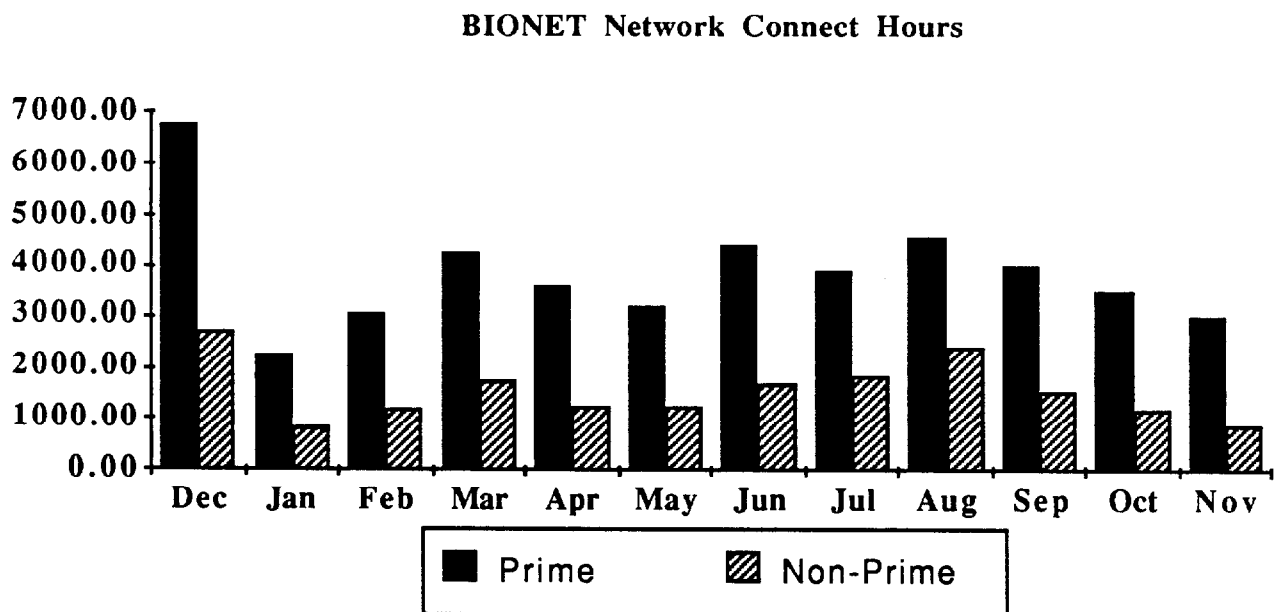
Figure III-5: Bionet's Total Use of the Dec-2065 12/87-11/88



Tabel III-9: BIONET Network Usage Connect Hours

	Prime	Non-Prime	Total
Dec	6779.69	2717.94	9497.63
Jan	2246.27	857.43	3103.70
Feb	3047.92	1203.13	4251.05
Mar	4279.67	1774.58	6054.25
Apr	3620.61	1277.32	4897.93
May	3210.61	1259.72	4470.33
Jun	4421.23	1724.26	6145.49
Jul	3928.71	1860.19	5788.90
Aug	4573.06	2395.25	6968.31
Sep	4009.63	1543.09	5552.72
Oct	3516.30	1228.15	4744.45
Nov	3019.07	902.89	3921.96
Total	46652.77	18743.95	65396.72

Figure III-9: Total Telenet and Compuserve Network Use, 12/87-11/88



3.1.5.4 Computer Software - Core Library

There was a new release (version 5.2) in August 1988 of the IntelliGenetics software suite that makes up the Core Software Library. This software is made available to the BIONET community immediately upon its formal release. Version 5.2 included the addition of new functionality to many of the existing programs and also included a completely new program, FINDSEQ, that simplified the retrieval of sequence data from the GenBank and PIR databases and also allowed for the performance of Boolean searches. FINDSEQ was considerably more user-friendly and powerful than the previously used FIND command, but was designed to handle searches that did not require the power of the IntelliGenetic's QUEST database pattern searching program. The next release of the IntelliGenetics software is scheduled for early 1989 and will be made available on BIONET. It will feature significant new improvements in speed and sensitivity for database similarity searches.

3.1.5.5 Computer Software - System Library

During the course of the year most new systems software was added to the new Sun network. This was documented above in the section entitled **Development work on the new Sun central computing facility**.

3.1.5.6 Computer Software - Contributed Library

Software contributed to BIONET is placed in the <CONTRIBUTED> directory on the DEC-2065, to which only the BIONET community has access. Recently BIONET has also implemented an "anonymous FTP" capability on the new Sun 3/280 file server. This allows users on other Internet computers to retrieve software from BIONET over the network through the use of the FTP or File Transfer Protocol. This method of retrieval is significantly faster than downloading files from BIONET to a remote PC and does not run up communications charges.

Major software packages produced by BIONET collaborators and implemented on BIONET with the aid of our staff have been summarized under *Collaborative Research*, above. Refer also to the software lending library catalog in *Appendix IV*.

3.1.5.7 Database Library

BIONET provides its users with a large number of different databases in support of molecular biology and molecular genetic research, the most popular being the Roberts' restriction enzyme database and the GenBank nucleic acid sequence database. We provide database updates in a timely manner to the community. Our sequence database releases in IntelliGenetics' file format have usually been 2-3 weeks after obtaining the tapes from NIH GenBank or the EMBL. The original database format files are made available on the DEC-2065 usually the same evening of the day on which the tapes arrive.

BIONET databases were discussed in detail in previous Annual Reports. New additions and changes to the databases on BIONET during this year are described above under *Data Contributors*. Here we document the use of the major databases on BIONET.

The DNA and protein sequence databases are used by BIONET scientists as a source of sequence data and for searching. Two major types of searches are performed. The main usage occurs when BIONET users search the database for similarities or homologies with sequences that they have determined. The second type of search involves use of the QUEST program to find interesting

consensus sequences that are known to have functional importance. This year our statistics on database searching have been dramatically impacted by the introduction of the FASTA-MAIL program. Since accounting software is not yet available on the Sun system we cannot provide accurate data for the number of individual GenBank, PIR, or SWISS-PROT database searches performed by users. Statistics can, however, be provided for the overall total number of searches. We know that the FASTA-MAIL interface on the DEC is utilized 950 times per month or 32 times per day on average. FASTA-MAIL is used to search each of the three databases just mentioned. BIONET has successfully promoted the use of FASTA-MAIL over the use of other database searching software to reduce the load on the DEC 2065 computer. In addition the IntelliGenetics' QUEST and IFIND programs are used 346 times per month for database searches.

Restriction Enzyme Database - Thanks to Dr. Richard Roberts, the chairman of the BIONET National Advisory Committee, we have established one of the most up-to-date lists of restriction enzymes available. Dr. Roberts maintains a restriction enzyme registry and distributes his updated lists in an electronic message to BIONET. BIONET incorporates these new lists into its program essentially immediately. These lists are used within the core program SEQ and PEP and are referenced 459 times per month.

VectorBanktm - Vectorbank is a collection of sequence data and restriction maps of important cloning vectors, viruses and phages that is maintained by the IntelliGenetics staff. This database is used by the CLONER program for manipulating restriction maps and simulating DNA cloning experiments. Release 4.2 of VectorBank in May saw the revision of the restriction map data files for all 146 vectors in the database. All single and double cutter sites using prototype enzymes from the latest restriction enzyme database were included in the map data. Individual Vectorbank sequence files are accessed up to 25 times per month.

KeyBanktm - KeyBank is a data bank of nucleic acid and protein consensus patterns collected from the literature by IntelliGenetics, Inc. KeyBank is produced in a format suitable for use with the QUEST pattern searching program of the IntelliGenetics Suite. Some of the patterns available include binding sites, active sites, allosteric sites, phosphorylation sites, cleavage sites, cap sites, chromosome modification sites, polymerase binding sites, insertion sites, regulatory sites, methylation sites, replication origins, satellite DNA, Z DNA, and zinc finger regions. The latest release (rel. 3.0, October 1988) has 1237 new patterns taken from more than 900 references. This part of KeyBank occupies more than 690,000 bytes of memory. In addition, KeyBank also has files of codon and restriction enzyme site patterns. The contents of the key files in the main part of KeyBank can be easily searched by using the many indices supplied with the database, such as the organism, keyword, or citation indices. Individual database files in KeyBank are utilized up to 10 times per month.

3.2 Highlights

The sections above describe in detail our accomplishments in the several components of the BIONET Resource. Considering that a very significant fraction of staff time during our fifth year was involved with planning, grant writing, and other obligations for our renewal, substantial progress was still made on the Resource. Here, in brief, are some of the most notable:

- The new computer network donated by Sun Microsystems is being prepared for direct access by the user community by the BIONET systems staff (Mr. Rob Liebschutz and

Mr. Eliot Lear). As of August 1988 BIONET released the FASTA-MAIL program. This provided our users on the DEC with electronic mail access to high-speed database searches on our Sun 3/280 computer. Database search times were reduced from hours to tens of minutes or less, and the average mid-day user load on the DEC 2065 **dropped by a factor of about three to five** since it was no longer being used for these compute-intensive tasks. FASTA-MAIL used the FASTA program obtained from Dr. William Pearson, and the mail server portion was developed by Mr. Liebschutz, Mr. Lear, and Mr. Spencer Yeh.

- The BIONET user community continued its vigorous growth, up 31% over last year for a total of 867 laboratories on the system. On the DEC 2065 total cpu usage increased by 65% and total connect hours by 36% as compared to last year.
- Dr. Jerzy Jurka, the BIONET Scientist, published important work in the area of repetitive DNA sequence analysis. In conjunction with this research, new functionality has been added to the Multiple Aligned Sequence Editor (MASE) by the BIONET applications programmer, Mr. Liang Jen Horng. This editor was originally developed by Dr. Jurka and Donald Faulkner at Dana Farber's Molecular Biology Computer Research Resource and then extended at BIONET over the past year. Work on the editor has also involved collaborators from the machine learning group at the University of California at Santa Cruz.
- Dr. Sunil Maulik has continued work on the RICH program which performs database searches for sequences of defined percent composition. Dr. Maulik has obtained some interesting preliminary results with the program. He has also been involved in developing a new Hypercardtm-based user interface for BIONET.
- The electronic communications network was significantly enhanced. The efforts of Dr. David Kristofferson led to the formation of the international BIOSCI bulletin board network. Besides BIONET in the U.S., other major BIOSCI distribution sites are situated in England, Ireland, Sweden, and Finland. Recipients of the bulletin boards from these sites are located in all parts of the globe. The bulletin boards are available to users on the ARPANET, BITNET, EARN, Usenet, NSFnet, and JANET. Users in any particular location need only post or receive messages from their closest site. Any posting at any center is automatically forwarded by the central BIOSCI sites to all other participants on the above-listed networks.
- Finally, the research conducted by BIONET's 867 laboratory groups was made significantly easier by a total revision of the BIONET documentation and the production of a new User Manual. This involved major efforts by BIONET staffer's Ms. Vickie Johncox, Mr. Spencer Yeh, and Ms. Kathryn Berg. The documentation was sent free of charge to all users on the system this past summer.

3.3 Administrative Changes

The following have been the administrative changes within BIONET during the past year. These have come about for reasons ranging from personnel shifts to additions and resignations. None of these changes has had a negative impact on the Resource itself, in particular, its "appearance" and availability to the community.

- Mr. Liang Jen Horng was hired in May 1988 as the new BIONET Applications Programmer to assist Dr. Jurka in his research efforts and to work on the BIONET contributed software. Mr. Horng has a M.S. in Computer Science from San Jose State University.
- Mr. Eliot Lear was hired in June by the Computer Facilities group at IntelliGenetics as an additional system programmer. Mr. Lear has a B.S. in Computer Science from Rutgers and has run a state-wide network of Sun Microsystems file servers and

workstations. Since joining IntelliGenetics he has devoted a large fraction of his time to bringing up the new Sun system for BIONET.

- Ms. Vickie Johncox, formerly a BIONET Scientific Consultant, left BIONET in June for another job as head of IntelliGenetics' training group. BIONET hired Dr. Karen Davis, formerly at the University of California, Santa Cruz, as her replacement. Dr. Davis started at the end of June. No disruption in the service occurred as Ms. Johncox's duties were handled smoothly by Mr. Spencer Yeh and Drs. Maulik and Kristofferson.
- In November, Ms. Kathryn Berg resigned as BIONET Administrator to pursue work elsewhere. Before her departure BIONET hired Ms. Cindy Eppard from IntelliGenetics to replace Ms. Berg. Ms. Eppard learned her new responsibilities rapidly, and there was no interruption in the establishment and processing of BIONET user accounts.

Despite the demands of their jobs, increased once again by a 31% growth in the size of the user community this year, the BIONET staff remains a highly dedicated group of individuals who are interested in their work and in promoting the goals of the resource. BIONET has received many testimonials to the quality of its staff and the support that they provide. Copies of a few of these testimonials are included in *Appendix VIII*.

3.4 Resource Advisory Committee and Allocation of Resources

The membership of BIONET's National Advisory Committee has not changed this past year. The NAC consists of:

- Dr. Richard J. Roberts, Ph.D., **Chairman**, Senior Staff Investigator, Molecular Biology, Cold Spring Harbor Laboratory
- Professor John Abelson, Ph.D., Department of Biology, California Institute of Technology
- Professor Alan Maxam, Ph.D., Dana Farber Cancer Institute, Harvard Medical School, Harvard University.
- Thomas Rindfleisch, M.S., Director, Knowledge Systems Laboratory, Department of Computer Science, Stanford University.
- Professor Irwin Kuntz, Ph.D., Department of Pharmaceutical Chemistry, University of California, San Francisco.
- Professor Charles Yanofsky, Ph.D., Department of Biological Sciences, Stanford University.
- Professor Eric Lander, Ph.D., Harvard Business School
- Professor Joshua Lederberg, M.D., Ph.D., President, The Rockefeller University.

The last regularly scheduled meeting occurred on November 13, 1987 in Mountain View. Due to the time occupied by preparations for the BIONET renewal the next NAC meeting has been postponed until the first part of 1989. When issues have arose during this last year we have consulted members of the NAC by phone, especially our chairman, Dr. Roberts, and Mr. Thomas Rindfleisch at Stanford.

Allocation of Resources. The Committee agrees with our methods for allocating the Resource. The DEC-2065 computer uses its windfall scheduler to allocate cpu time to the various categories of users and overhead, as described under *Resource Facilities*. The cpu time is distributed on a first-come, first-served basis. This method has been very successful. Considerably more than 50% of CPU time (BIONET's original allocation) has been delivered to BIONET scientists (see *Resource*

Facilities, above). BIONET is now routinely using almost 90% of the DEC 2065 plus significant resources on the new Sun computer system.

We continue to request that the community not have more than one person per PI group using BIONET at the same time during prime time. The community continues to do an excellent job in complying with this policy.

We continue to allocate additional disk space to PI groups involved in managing large sequencing projects or extensive databases of sequences. We do this on an *ad hoc* basis upon requests by investigators. Our archive and retrieval system is working smoothly for archival storage and prompt retrieval (one to two days) of important files.

3.5 Dissemination of Information of Resource's Capabilities

We discuss two areas related to dissemination of information about the Resource that we have pursued this grant year. The first is interactions with the scientific community through participation at conferences and seminars. The second is use of the electronic mail and bulletin board facilities of the Resource itself to keep scientists worldwide aware of changes and improvements.

3.5.1 Community Interactions and Awareness

We have used two methods this year to inform the community about BIONET and to solicit applications for access to the Resource. The first method has been the presentation of invited seminars and participation at major conferences where we have presented lectures and/or have had booths at exhibitions. These efforts are summarized above under *BIONET Training Program*. At these conferences, we have distributed the standard applications packets to scientists, after demonstrating to them the capabilities of the Resource.

The second method has been through the publication of journal articles which describe BIONET. This year an article detailing new developments at BIONET appeared in vol. 16 no. 5 of *Nucleic Acids Research*. A preprint of this article was in last year's Annual Report. An upcoming article will appear in *Protein Sequence and Data Analysis* and is listed on the BRTP Training Form under *Scientific Subprojects*. A copy of this article is available in *Appendix II*.

3.5.2 Electronic Communications

The electronic communication facilities of BIONET provide another important way to disseminate information about the Resource. In addition, electronic mail and bulletin boards provide a mechanism for scientific and technical interchanges among members of the community. With the dissemination of BIONET bulletins to investigators outside of the Resource (via ARPANET, BITNET and USENET), information about BIONET is being distributed electronically worldwide. Information on the types of electronic mail communications with BIONET was summarized previously in the discussion of the *Collaborative Research* component of the Resource.

3.6 IMPORTANT Suggestions and Comments

For the BIONET staff the main "highlight" of this year was the renewal process of the BIONET grant. We have not commented on this topic in this report because some issues are still under negotiation with the NIH and are not yet for public disclosure. We do, however, have the following general comments.

The regulations of the BRTP program at the Division of Research Resources require a Resource such as BIONET to excel in several different areas: Technological Research, Collaborative Research, Service, Training, and Dissemination. Given the size of the BIONET user community and the demands that it makes on our staff, it has always been a challenge to cover all of these areas adequately. We have excelled in the areas of Service, Training, and Dissemination because, until the addition of Dr. Jurka, that was where our talents lay and the sheer din of the user community would have been deafening had we neglected the Service to devote more of our limited resources to research. We feel that the real justification of this resource has been the quality of the research done **BY OUR HUNDREDS OF USERS** and that any efforts on our behalf would always be dwarfed by comparison.

Nonetheless, DRR regulations make it possible for a committee of ten people **who are not even users of the Resource** to come in and possibly scuttle the entire operation because of deficiencies in just **one** of the five mandated areas. Furthermore it is possible that such an event can occur without **ANY** input from the almost 3000 users of the resource! The fact that such an action could lead to the disruption of research and electronic communications for almost 900 laboratories around the world does not seem to be a perturbing factor in the renewal decision. **This surely must be one of the more amazing possibilities in the history of modern science!**

If this year's report demonstrates anything, it shows that **BIONET works and works successfully**, and that it has continued to make significant improvements over the last several years. We have the largest user community of any similar resource in the world and it continues to grow, even attracting users from locations that have competing facilities. Few, if any, other related organizations can lay claim to the "resourcefulness" that BIONET has shown in successfully pursuing its proposal to Sun Microsystems which resulted in the grant of \$150,000 in new hardware.

We suggest that reason finally prevail in this process, and that the NIH recognize, preserve, and expand this valuable resource. DRR regulations should be reviewed and revised so that organizations that are primarily devoted to Service need not excel in every single BRTP category in order to survive.

To date the user community is only dimly aware of the forces impinging upon BIONET. Because of the current negotiations with the NIH it has not been appropriate to involve them. It would be extremely interesting to see their reaction if events lead to the termination of the BIONET Resource, but, if reason prevails, this disastrous event will not come to pass, and the users will continue to invest their energies productively in their research.

I. BIONET Research Publications

Copies of six BIONET research publications and abstracts are included in this section.

A fundamental division in the *Alu* family of repeated sequences

(evolution/*Alu* subfamilies/secondary structure/CpG dinucleotide)

JERZY JURKA*[†] AND TEMPLE SMITH[‡]

*Bionet, 700 East El Camino Real, Mountain View, CA 94040; and [‡]Dana-Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street, Boston, MA 02115

Communicated by Roy J. Britten, March 10, 1988

ABSTRACT The *Alu* family of repeated sequences from the human genome contains two distinct subfamilies. This division is based on different base preferences in a number of diagnostic sequence positions. One subfamily of the sequences, referred to as the *Alu-J* subfamily, is very similar to 7SL RNA in these positions. The other subfamily, *Alu-S*, can be divided further into well-defined branches of sequences. These findings revise the previous picture of the *Alu* family and expose their complex evolutionary dynamics. They reveal sequence variations of potential importance for the proliferation of *Alu* repeats and relate them to their structural features. In addition, they open the possibility of using different types of *Alu* sequences as natural markers for studying genetic rearrangements in the genome.

A typical human *Alu* family member is a sequence ≈ 300 base pairs long and consists of two similar but not identical subunits, *Alu-left* and *Alu-right*, connected by an adenine-rich linker. Both halves of *Alu* elements are related to the 7SL RNA (1). Although *Alu* sequences are the most abundant among middle repetitive elements in the human genome, their biological role remains unclear (2). In this paper, we report on the presence of at least four different types of *Alu* sequences, which probably originated at different times in the history of primates.

METHODS

A set of 125 complete or nearly complete human *Alu* sequences were extracted from the GenBank DNA sequence data base.[§] The list of the GenBank loci used, positions of the extracted sequences, and other specifications are given in the legend of Table 1.[¶] The statistical analysis described below is based exclusively on pairwise comparisons of each *Alu* sequence with the consensus sequence (see Fig. 1), using the computer algorithm of Smith and Waterman (3). The overall consensus sequence in Fig. 1 was derived from our data and is slightly different from the one recently published (2). The differences are exclusively within CpG doublets, which are known to be variable in *Alu* repeats (4). Taking pairwise comparisons as a starting point, the multiple alignment of the analyzed set of sequences has been done by hand with a specialized sequence editor (5). To detect sequence positions with different base preferences (diagnostic positions), we used "column-correlation" function incorporated in the sequence editor (5). This function was originally designed to perform automatic searches for compensatory mutations.

RESULTS

During a search for compensatory mutations in the multiply aligned set of 125 *Alu* sequences, we noted an unusually high

proportion of correlated base occurrences in at least 15 sequence positions. These positions are referred to as diagnostic positions and are listed in column 1 of Table 1 (the position numbers are the same as in Fig. 1). The observed correlations in the diagnostic positions reflect different base occurrences in different *Alu* subfamilies. It is shown below that the most predominant bases in the 15 diagnostic positions belong to only one of the two basic types of *Alu* sequences present in the analyzed set.

To segregate the most predominant type from the remaining *Alu* sequences, we have used computer alignment (3) of each *Alu* element with the *Alu* consensus from Fig. 1. The average overall similarity between the 125 *Alu* sequences and the *Alu* consensus is 83.88% with a SD of 5.63% (gaps counted as single mismatches). These numbers are slightly different if gaps are excluded from the analysis (see Table 3). Given the overall similarity, we assume that the probability of matching between any *Alu* sequence and its consensus in a randomly chosen aligned position equals 0.83. Any *Alu* sequence similar 40% or less to the consensus sequence in the 15 diagnostic positions has been defined as an *Alu-J* element. The probability of only 6 matches or less in 15 randomly chosen aligned positions can be calculated from the binomial distribution and is <0.001 . Following the statistical definition, we have found 31 *Alu-J* sequences in the analyzed set of 125 sequences. The remaining 94 sequences are referred to as *Alu-S* sequences. The 3:1 ratio of S/J *Alu* sequences explains why the overall consensus sequences and *Alu-S* consensus sequence overlap. We have found no sequences matching seven or eight diagnostic consensus positions, which suggests that the distinction between J and S sequence types is quite sharp with few or no intermediate forms. As shown in Table 1, in the diagnostic positions the J subfamily maintains consistently different bases from those in the S subfamily. The difference in base preferences between J and S subfamilies is most evident at positions 94, 204, and 275 (Table 1). For example, G-204 is present in 29 of 31 *Alu-J* sequences and in only 1 of 94 *Alu-S* sequences. Similarly, G-94 and C-275 are powerful diagnostic indicators that can be used for preliminary "by eye" identification of *Alu-J* elements.

As illustrated in Table 1 the most frequent bases in the J subfamily are identical with those in 7SL RNA in 14 of 15 diagnostic positions. Furthermore, the differences between J and S *Alu* elements correlate with differences in the adenine-rich linker connecting the left and right halves of the *Alu* dimer (positions 121-133 of the consensus sequence in Fig. 1; data not shown). The triplet TAC in the middle of the linker is present in $\approx 80\%$ of *Alu-S* compared to only 20% of the *Alu-J* sequences. It is not certain if the homologous TAC triplet was ever present in many *Alu-J* sequences since their

[†]To whom reprint requests should be addressed.

[§]EMBL/GenBank Genetic Sequence Database (1987) GenBank (IntelliGenetics, Mountain View, CA), Tape Release 46.0.

[¶]The *Alu* sequences used in this study are available on the Bionet computer in the file <jurka>human-alu.seq. The data can also be obtained by electronic mail from jurka@bionet-20.arpa.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Diagnostic base differences between major subfamilies of the *Alu* family

Consensus position	<i>Alu</i> subfamily	Frequency of					Base in 7SL DNA	Consensus position	<i>Alu</i> subfamily	Frequency of					Base in 7SL DNA
		T	C	A	G	(-)				T	C	A	G	(-)	
57 (C)	J	6	3	20	0	2	A	163 (A)	J	1	2	3	25	0	G
	S	36	47	9	2	0			S	2	1	55	36	0	
63 (A)	J	2	0	12	16	1	G	194 (A)	J	1	0	6	22	2	G
	S	1	2	87	4	0			S	1	1	91	1	0	
65 (C)	J	30	1	0	0	0	T	204 (A)	J	0	0	1	29	1	G
	S	20	41	1	1	37			S	0	0	89	1	2	
70 (G)	J	2	21	0	7	1	T	208 (G)	J	0	0	20	11	0	A
	S	0	2	3	89	0			S	2	3	29	59	1	
71 (T)	J	5	23	0	2	1	C	220 (T)	J	6	22	1	1	1	C
	S	90	4	0	0	0			S	80	10	0	1	3	
94 (C)	J	0	0	2	29	0	G	233 (A)	J	22	2	4	0	3	T
	S	4	87	1	1	1			S	0	0	90	3	1	
101 (G)	J	2	0	19	10	0	A	275 (T)	J	1	28	1	0	1	C
	S	0	0	9	85	0			S	85	2	1	2	4	
106 (A)	J	0	0	13	18	0	G								
	S	0	0	93	1	0									

Consensus positions are taken from Fig. 1. (-), Alignment gaps. Loci names and 5' → 3' positions of *Alu*-J and *Alu*-S sequences are listed below as they appear in GenBank (release 46.0).⁵ *Alu* repeats complementary to the consensus sequence are listed in 3' → 5' order. Positions preceded by b and c indicate b and c branches of *Alu*-S sequences, respectively, as defined in Table 2 and in the text. *Alu*-J: HUMACHRA7(1580-1295); HUMADAG(4907-5201, 24773-24495, 31460-31747); HUMAPOCII(1982-2235); HUMAPOE4(2562-2849); HUMBLYM1(266-560); HUMERPA(1810-2100); HUMFIXG(24172-24465); HUMFOLS(1577-1847); HUMIFNB3(2663-2405, 13213-13489); HUMIL2R8(1209-1486); HUMLDLR(4193-4485); HUMPOMC2(26-340); HUMPOMC6(26-303); HUMRSAOLD(498-790); HUMRSKPA1(24-291); HUMTBB5(2922-2627, 2949-3239, 5611-5885); HUMTHBNB(3593-3883); HUMTPA(7512-7227, 8862-9165, 10801-10513, 16794-17114, 18878-19167, 20944-21250, 22262-22536, 26941-27228); M13121(1141-856). *Alu*-S: HUMA1ATP(4932-5219); HUMADAG(1672-1369, 2357-2642; c: 5606-5893, 8000-7720, 8484-8193, 13452-13741, 15386-15096, 15806-16094, 17224-16933, 18414-18706, 19900-19613, 22527-22812, 25453-25163, 27269-26979; c: 28032-28320); HUMAGG(b: 1391-1106); HUMALBG(3287-3576, 6046-5759); HUMANFA(c: 1340-1621; b: 1630-1919); HUMAPOAI1(3291-3585, 6421-6709); HUMAPOAI2(2571-2860); HUMAPOCII(2254-2542); HUMAPOE4(636-352, 2427-2138, 5049-4773); HUMCIA21(347-60); HUMCIA23(330-45); HUMCIAIN1(992-1285); HUMFIXG(7298-7595; c: 31537-31801, 35947-36248); HUMFOLS(1284-989); HUMGAST2(187-477); HUMGHV(2506-2248); HUMHBA4(2060-1773, 4297-4585, 8548-8836); HUMHBBRT(482-190, 1260-1548); HUMIFNB3(4648-4363, 7265-7545; c: 8975-8688); HUMINS2(69-357); HUMLDLVS(291-8); HUMLDLR(b: 3715-4011); HUMMHDC3B(b: 3712-3424); HUMMHDRB3(b: 2838-3124; c: 4063-4345); HUMMYCRT(c: 3143-2876); HUMNGFB(c: 5259-5544); HUMPOMC1(1392-1102, 7099-6803); HUMPOMC1(333-47); HUMRSA1(c: 508-803); HUMRSA27(1-251); HUMRSA16(b: 168-451); HUMRSAB11(1-269); HUMRSAB13(11-295); HUMRSAB19(1-241); HUMRSAB2(1-288); HUMRSAB6(1-256); HUMRSAB8(1-265); HUMRSAP3(b: 897-1186); HUMRSKA1(21-347); HUMSLJT1(568-280); HUMTBB5(3289-3573, 4115-3849; b: 5241-4953; b: 6799-6516); HUMTBBM40(c: 1828-2113); HUMTHBNB(1165-874, 3418-3110); HUMTPA(5960-5671, 6746-6483, 739-1022, 10066-10355, 12986-12700; b: 17170-17455, 21279-21567, 21940-21651, 25619-25905, 26522-26811, 27879-28149; b: 28803-29090, 32922-33210, 34234-34503); HUMUG2PD(c: 546-260, 1685-1396); M11591(b: 1404-1115); M12036(637-362); M12929(592-302).

linkers vary substantially in both their size and the primary sequence.

Following the same approach, the S subfamily of the *Alu* family has been found to contain other types of *Alu* sequences. Unlike the J/S division, the intra-S division is more difficult to define statistically since the number of simultaneous differences between subsets of *Alu*-S sequences appears to be smaller and there is a number of intermediate sequences virtually absent from the J-S junction. Therefore, we first define the most distinct "b" branch of the S subfamily as containing sequences that match 3 or fewer of the 11 diagnostic positions listed in Table 2 and in Fig. 1. There are 12 such elements in the analyzed set of 94 *Alu*-S sequences. The average overall similarity between each analyzed *Alu*-S sequence and the consensus sequence is 86.59 if every gap is counted as a single mismatch (Table 3). Based on this number, we assume the probability of matching the aligned consensus sequence at a randomly chosen position to be 0.86. As calculated from the binomial distribution, the probability of matching 3 or fewer of the randomly chosen aligned positions is 10^{-4} . The probability of matching exactly 4 and 5 positions equals 1.9×10^{-4} and 1.63×10^{-3} , respectively. We have found 5 sequences matching 4, and 6 matching 5 diagnostic positions in the analyzed set of 94 *Alu*-S sequences. These 11 sequences are arbitrarily defined as a "c" branch of the S subfamily. After segregation of the b and c branches, the remainder of the *Alu*-S subfamily is referred to as an "a-branch." Preliminary analysis of 71 *Alu*

elements from this branch revealed the presence of 16 sequences containing simultaneously thymine at position 244 and adenine at position 272, as opposed to C-244 and G-272 in the remaining 55 *Alu* sequences. In addition, 14 of the above 16 sequences contain an extra adenine in position 264. This suggests that the *Alu*-a branch may contain at least two different types of *Alu* sequences and it can tentatively be replaced by "d" and "e" branches containing 16 and 55 sequences, respectively.

As illustrated in Table 2 and Fig. 1, the base preferences are quite similar between *Alu*-b and *Alu*-c sequences up to position 88. Further on, *Alu*-c remain similar to *Alu*-a with the exception of guanine at position 163. Therefore, the c branch can be viewed as an intermediate between the a and b branches of the S subfamily. A unique feature of the c sequences may be the presence of adenine at position 74. Of the 125 *Alu* sequences only 8 contain A-74 of which 7 belong to the *Alu*-c branch defined above. The eighth *Alu* sequence containing adenine at position 74 (HUMPOMC1) has all the *Alu*-c features listed in Fig. 1: deletion at 64 and 65, A-78, T-88, and G-163. Therefore, it can also be considered as an *Alu*-c sequence.

Based on the analysis of phylogenetic trees, other authors (6) have recently identified the *Alu*-b branch as a "subfamily of the *Alu* family." The authors have pointed out differences between the *Alu* consensus and the *Alu*-b sequences at positions listed in Table 2 as well as in Fig. 1 and at other less characteristic positions not included in our analysis.

	1	15	16	30	31	45
7SL	-GCCGGGCGCGGTGG	CGCGTGCCTGTAGTC	CCAGCTACT-CCGGGAG			
Alu-cons	GGCCGGGCGCGGTGG	CTCACGCCTGTAATC	CCAGC-ACTTTGGGAG			
	46	60	61	75	76	90
7SL	GCTGAGGCTGGAGGA	TCGCTTGAGTCCAGG	AGTTC...CCAGCC			
Alu-J		a	g t CC			
Alu cons	GCCGAGGCGGGCGGA	TCACCTGAGGTCAGG	AGTTCGAGACCAGCC			
Alu-c		--	(A)	A		T
Alu-b		--		A		T
	91	105	106	120	121	135
7SL	TGGGCAACATAGCGA	GACCCCGTCTCT				
Alu-J	G	a	g			
Alu cons	TGGCCAACATGGTGA	AACCCCGTCTCTACT	AAAAATACAAAATT			
Alu-c						
Alu-b	T	C				
	136	150	151	165	166	180
7SL	-GCCGGGCGCGGTGG	CGCGTGCCTGTAGTC	CCAGCTACTCGGGAG			
Alu-J			g			
Alu cons	AGCCGGGCGGTGGTGG	CGCGCGCCTGTAATC	CCAGCTACTCGGGAG			
Alu-c			G			
Alu-b		G	G			
	181	195	196	210	211	225
7SL	GCTGAGGCTGGAGGA	TCGCTTGAGTCCAGG	AGTTCTGGGCTGTAG			
Alu-J		G	G a			C
Alu cons	GCTGAGGCAGGAGAA	TCGCTTGAACCCGGG	AGGCGGAGGTTGCAG			
Alu-c						
Alu-b		G	R			C
	226	240	241	255	256	270
7SL	TGCGCCTGTGA...G	CCACTGCACTCCAGC	CTGGGCAACATAGCG			
Alu-J		T				
Alu cons	TGAGCC-GAGATCGCG	CCACTGCACTCCAGC	CTGGGCGACAGAGCG			
	271	285				
7SL	AGACCCCGTCTCT					
Alu-J		C				
Alu cons	AGACTCCGTCTCAAA	AAAAA				

FIG. 1. Consensus sequence for 125 *Alu* sequences and the homologous regions of human 7SL DNA. Major and minor characteristic bases for other types of *Alu* sequences are printed in capital and lowercase letters, respectively, and correlate with the analysis in Table 1. Dots indicate sequence regions absent from 7SL DNA but present in the *Alu* family. The remaining 7SL-specific sequences are not shown. Dashes under positions 64 and 65 indicate bases missing in *Alu-b* and *Alu-c* sequences. Additional characteristic positions not listed in Table 2 are put in parentheses.

The diagnostic position 78 (Table 2) is in the middle of the stretch 77–79, which can pair with base 87–89 containing another diagnostic position 88. Bases 77–79 are within the polymerase III promoter region (bases 74–86 in Fig. 1). Correlation between occurrences of complementary bases at positions 78 and 88 suggests the possibility of a weak secondary interaction in this region. Another potential for secondary interaction, already proposed for 7SL RNA (7, 8), exists between complementary bases 69–75 and 89–95. This region includes 3 of the 15 positions distinguishing between the J and S subfamilies and the complementarity is conserved throughout the *Alu* family. The only A-C mispairing has been found in this region in the *Alu-c* sequences. The role of the above hypothetical structures is not clear, although their location suggests involvement in *Alu* transcription. There is also a possibility of a secondary interaction between bases 244 and 245 and bases 271 and 272 that includes bases at positions diagnostic for putative d and e branches of the *Alu* family discussed above.

Table 3 indicates that the average overall similarity between *Alu-J* and the *Alu* consensus sequence in nondiagnostic positions is lower than the average similarity between *Alu-S* and the *Alu* consensus. This indicates that on average *Alu-J* sequences are more diverse than *Alu-S* sequences. By

t test, one can find that differences between *Alu-J*/consensus and *Alu-S*/consensus similarities are statistically significant ($P < 0.001$). The conclusion holds true even if the general *Alu* consensus is replaced by the *Alu-J* consensus (data not shown). There is also a significant difference ($P < 0.001$) between analogous numbers for a and b subdivisions of the *Alu* sequences. The differences between *Alu-b* and *Alu-c* sequences are marginally significant ($P < 0.05$), and analogous differences between *Alu-a* and *Alu-c* are insignificant.

As pointed out before (4), CpG doublets undergo rapid mutations in *Alu* sequences. This may result from a deamination of methylated cytosine (for a review, see ref. 9). Average CpG content is lowest in the J subfamily (3.84 ± 2.01) as compared to analogous numbers for *Alu-a* (7.75 ± 2.95), *Alu-b* (16.08 ± 5.01), and *Alu-c* (9.54 ± 3.75) branches of the S subfamily. Significance levels for the differences in the CpG content are virtually identical to those for the similarity differences discussed in the preceding paragraph.

DISCUSSION

Given the similarity between *Alu-J* and 7SL RNA sequences in the diagnostic positions, the large intra-subfamily diversity and the low CpG content, we find the J sequences to be good

Table 2. Base preferences in the S subfamily branches

Consensus position	Branches of <i>Alu</i>	Frequency of				(-)
		T	C	A	G	
65 (C)	a	20	41	1	1	8
	c	0	0	0	0	11
	b	0	0	0	0	12
66 (T)	a	62	3	3	0	3
	c	4	0	0	0	7
	b	4	0	0	0	7
78 (T)	a	67	0	4	0	0
	c	1	0	9	1	0
	b	0	0	12	0	0
88 (G)	a	2	1	2	65	1
	c	9	0	1	1	0
	b	11	0	1	0	0
95 (C)	a	2	68	1	0	0
	c	2	9	0	0	0
	b	12	0	0	0	0
100 (T)	a	66	1	2	1	1
	c	7	4	0	0	0
	b	1	10	1	0	1
153 (C)	a	10	35	1	24	1
	c	5	1	0	5	0
	b	0	1	0	11	0
163 (A)	a	1	0	53	17	0
	c	1	1	1	8	0
	b	0	0	1	11	0
197 (C)	a	15	50	2	4	0
	c	3	6	1	1	0
	b	0	0	0	12	0
200 (T)	a	65	3	2	1	0
	c	10	0	1	0	0
	b	1	0	4	7	0
219 (G)	a	0	1	2	64	0
	c	0	1	0	10	0
	b	0	11	0	0	0

(-), Alignment gaps.

candidates for the early *Alu* elements derived from the 7SL RNA (1). The base differences in the diagnostic positions and the linker regions may be important for understanding how this transformation occurred and are good targets for experimental analysis. On the other hand, the least diverse *Alu*-b sequences can be viewed as a relatively young branch of the *Alu* family. There are three published examples of *Alu* sequences that are believed to be inserted relatively recently on the evolutionary time scale: in the α -satellite DNA of African green monkey (10), in the gorilla β -globin gene cluster (11), and at the *MLV*-2 locus of human cell lymphoma (12). All these *Alu* sequences belong to the b branch defined above.

While this paper was in review, other authors (13) reported on a subdivision of the *Alu* family into three different subfamilies corresponding to our J subfamily and two branches (a and b) of the S subfamily. These two branches, as well as the branch c, are virtually equally different from the J subfamily of *Alu* sequences and similar to each other in the

Table 3. Average overall similarities with the *Alu* consensus sequence

<i>Alu</i> type	Gaps as mismatches		Gaps excluded		Total
	Mean	SD	Mean	SD	
All	83.88	5.63	86.39	4.38	125
J	79.20	4.35	82.83	2.27	31
S	86.59	3.39	88.75	1.98	94
a	86.52	3.26	88.59	1.79	71
c + b	89.20	4.14	91.15	3.08	23
c	87.04	4.25	89.45	2.87	11
b	91.22	2.92	92.71	2.45	12

Sequence alignments have been made by using the computer algorithm (2). The diagnostic positions have been excluded from similarity calculations.

diagnostic positions from Table 1. Therefore, we consider them as members of the S subfamily. The authors draw their conclusions from analysis of pairwise difference distribution among *Alu* sequences involving both the diagnostic differences discussed in this paper and a mutational noise. Our analysis is based on multiple sequence comparisons, which permits more rigorous distinction between diagnostic and background differences. With this level of resolution we are able to classify each *Alu* sequence individually. This, and the analysis of the CpG content discussed in the accompanying paper (14), opens a way to date the invasion of individual genes by different types of *Alu* sequences and of genetic rearrangements associated with this process.

We thank Donald Faulkner for professional computer assistance and Roy Britten, Douglas Brutlag, Terry Friedemann, David Kristoferson, and Randall Smith for critical and useful comments on the manuscript.

- Ullu, E. & Tschudi, C. (1984) *Nature (London)* **312**, 171-172.
- Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S. & Matsubara, K. (1987) *Gene* **53**, 1-10.
- Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **145**, 195-197.
- Bains, W. (1986) *J. Mol. Evol.* **23**, 189-199.
- Faulkner, D. V. & Jurka, J. (1988) *Trends Biochem. Sci.*, in press.
- Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. & Deininger, P. (1987) *Mol. Biol. Evol.* **4**, 19-29.
- Gundelfinger, E. D., Di Carlo, M., Zopf, D. & Melli, M. (1984) *EMBO J.* **3**, 2325-2332.
- Zwieb, K. (1985) *Nucleic Acids Res.* **13**, 6105-6124.
- Bird, A. P. (1987) *Trends Genet.* **3**, 342-347.
- Grimaldi, G. & Singer, M. F. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1497-1500.
- Trabuchet, G., Chebloune, Y., Savatier, P., Laucher, J., Faure, C., Verdier, G. & Nigon, V. M. (1987) *J. Mol. Evol.* **25**, 288-291.
- Economou-Pachnis, A. & Tschlis, P. N. (1985) *Nucleic Acids Res.* **13**, 8379-8387.
- Willard, C., Nguyen, H. T. & Schmid, C. W. (1987) *J. Mol. Evol.* **26**, 180-186.
- Britten, R. J., Baron, W. F., Stout, D. & Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4770-4774.

Multiple aligned sequence editor (MASE)

Donald V. Faulkner and Jerzy Jurka

Cognitive capacities of the human brain can not, so far, be matched by computers. Even well optimized computer programs have limited flexibility in addressing the variety of problems associated with sequence analysis. Hence, we were motivated to design a Multiple Aligned Sequence Editor (MASE) which combines manual sequence manipulations with standard computer analysis. An earlier article in *TIBS* described the adaptation of standard word processor software for a similar purpose¹.

MASE can be used for editing any set of sequences. The total number and size of sequences that can be edited simultaneously depends on the computer (typically, the total number of characters should not exceed 1×10^6). Sequences can be displayed on the screen in two windows, each of which can be moved either independently or concurrently. An example of the sequence display is shown in Fig. 1. The window sizes depend upon the type of terminal used. A standard VT100 ter-

minal can display 21 sequences per screen, with 25 characters in each window. Any terminal size and type which supports full screen editing can be used by writing a UNIX termcap entry. MASE can be run on computers with the Berkeley UNIX operating system.

Intrinsic functions

MASE provides over 80 intrinsic functions, each of which can be selected with few key strokes from a menu by calling the function COMMAND-MODE(:). Individual functions can also be fixed to different keys on the keyboard using the BIND function, and this assignment can be stored in a separate file and loaded whenever the sequence editor is used. A short definition of each intrinsic function can be called up using question-mark key. Full on-screen HELP and a tutorial are also available. The intrinsic MASE functions can be divided into the following basic groups: (1) moving cursors and searching for patterns; (2) modifying the sequence data; (3) changing the sequence display without affecting input/output files; (4) window manipulations; (5) sequence analysis; (6) modifying MASE behaviour; (7) mis-

cellaneous, and (8) generation of formatted output. In this short article we will outline only some of the capabilities of these functions.

Primary modifications of the sequence data involve insertions/deletions of alignment gaps in individual sequences or in the whole set. This permits the alignment for maximum similarity. One can begin with totally unaligned sequences or use an output from any sequence alignment program as a starting point. However, the format of the pre-aligned sequences must be as described below under the 'Sequence data files' and in the MASE manual. The on-screen alignment can be facilitated by changing the sequence display, for example by highlighting conserved sequence patterns, hydrophobic/hydrophilic residues, etc. (see Fig. 2). One can also easily emphasize differences between aligned sequences. The sequences can be rearranged arbitrarily to facilitate direct by-eye comparisons. Furthermore, the window display permits any two columns of characters be placed next to each other (e.g. columns 45 and 71 in Fig. 1).

Sequence analysis involves on-screen computation of consensus sequence, identity matrix, column com-

D. V. Faulkner is at the Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA. and J. Jurka is at Bionet, 700 East El Camino Real, Mountain View, CA 94040, USA.

Numbr	Locus Name	1	45	71	115
1	K1HUAG	-DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
2	K1HUAU	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
3	K1HUBI	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
4	K1HUAR	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
5	K1HUDE	B-IZMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	B-IZMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	B-IZMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	B-IZMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
6	K1HUEU	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
7	K1HUGL	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
8	K1HUHU	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
9	K1HUKA	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
10	K1HUKU	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
11	K1HULY	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
12	K1HUOU	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
13	K1HURE	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
14	K1HURY	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP	DIQMTQSPSSLSASVGDRTVITCRASQDINHYLNMYQQKPKAP
Offset from start		73, ignoring gaps	72		

Numbr	Locus Name	1	10	11	20
1	K1HUAG	DIQM	QEP	EL	ASVGDRV
2	K1HUAU	DIQM	QEP	LA	ASVGDRV
3	K1HUBI	DIQM	QEP	LA	ASVGDRV
4	K1HUAR	DIQM	QEP	LA	ASVGDRV
5	K1HUDE	B-IZM	QEP	EL	ASVGDRV
6	K1HUEU	DIQM	QEP	LA	ASVGDRV
7	K1HUGL	DIQM	QEP	LA	ASVGDRV
8	K1HUHU	DIQM	QEP	LA	ASVGDRV
9	K1HUKA	DI-QM	QEP	TL	ASVGDRV
10	K1HUKU	DIQM	QEP	QP	ASVGDRV
11	K1HULY	DIQM	QEP	LA	ASVGDRV
12	K1HUQU	DIQM	QEP	LA	ASVGBRV
Offset from start		1, ignoring gaps		0	

Fig. 2. A sequence display from MASE with T and S highlighted.

position and search for compensatory mutations using COLUMN-CORRELATION function. COLUMN-COMPOSITION permits analysis of base/amino acid frequencies at homologous sequence positions. COLUMN-CORRELATION lists the total number of simultaneous and non-simultaneous base differences between the aligned sequences in any two sequence positions. Each sequence can be compared either to a reference sequence (order independent), or to its nearest neighbor (order dependent). The probability of simultaneous and non-simultaneous base variations in the whole set is evaluated using the binominal formula. The COLUMN-CORRELATION function has

recently been used for secondary structure prediction and classification of human Alu sequences².

Sequence data files

Each input file contains one or more sequences. A sequence can contain an indefinite number of comment lines, each having a semicolon in the first column. The first line without semicolon contains the locus name which can not exceed 20 characters, and this is followed by the sequence information lines each not exceeding 95 characters in length. After insertion of alignment gaps the modified sequence set is stored in the same input format. Aligned output file can be created by a separate

OUTPUT-ALIGNED function. Line length, number of lines per page, positions of vertical and horizontal gaps, all depend upon the chosen output format.

Program availability

By the end of this year MASE will be made available on-line for the Bionet community. Other non-profit users can obtain the source code for editor written in C and MASE manuals from: Susan Russo, MBCRR, Dana-Farber Cancer Institute, 44 Binney Street, Boston MA 02115, USA. Tel. (617) 732-3746.

Acknowledgements

This software was developed under auspices of Molecular Biology Computer Research Resource which is supported by NIH grant RR02275-03. A portion of the support came from the NIH grant U41-RR-01685-05 for the Bionet resource. We thank Dr Temple Smith for helpful suggestions and strong encouragement for this project.

References

- 1 Boswell, R. B. (1987) *Trends Biochem. Sci.* 12, 279-280
- 2 Jurka, J. and Smith, T. F. (1988) *Proc. Natl Acad. Sci. USA* (in press)

Small cytoplasmic *Ro* RNA pseudogene and an *Alu* repeat in the human α -1 globin geneJerzy Jurka, Temple F. Smith¹ and Damian Labuda²

Bionet, 700 East El Camino Real, Mountain View, CA 94040, ¹Dana-Farber Cancer Institute, Harvard University, 44 Binney Street, Boston, MA 02115, USA and ²Medical Genetics, Hopital Sainte-Justine, 3175 Cote Sainte-Catherine, Montreal, Quebec H3T 1C5, Canada
Submitted November 6, 1987

The 5'-end of the previously studied *Alu* repeat from the α 1-globin gene (1) is flanked by a sequence 80% similar to one of the full length human small cytoplasmic *Ro* RNAs (Fig. 1a), denoted as HY3 (2). This is the first known example of a pseudogene for the *Ro* scRNA. Only a few such pseudogenes are expected to exist in the human genome (2). The pseudogene location next to the *Alu* sequence may suggest physical interactions between HY3-like RNA and the *Alu* RNA prior to the reverse transcription. The 3'-flanking region of the previously studied full size *Alu* repeat is another unreported *Alu* sequence truncated at the *Eco* RI restriction site (Fig. 1b). *In vitro* transcription of the region analysed (1) gave four RNA fragments. One of them, 86 nt long, is synthesized from the short class III transcriptional unit located on the 5'-side of the *Alu* repeat (3). This location coincides with the location of the HY3-like DNA sequence.

promoter?	GTGG-CNNAGTGG	
HY3	GGCTGGTCCGAGTGCAGTGGTGTTCACAACTAATTGATCACAACCAGTTA	50
	***** ***** * * ***** ***** *****	
3'- α 1	GGCTGGTTGGAGTGCAGCGCTTTTACAATTAATTGATCAGAACCAGTTA	52
		(a)
HY3	CAGATTTCTTTGTTCTCTCCACTCCCACTGCTTCACTTGACT-AGCCTTT	101
	* **** * ***** ***** *****	
3'- α 1	TAAATTTATCATTCTCTCCACTCCTGCTGCTTCAGTTGACTAAGCCTAA	104
promoter	GTGGCANNAGTGG	
Alu	GGCCGGGCGCGGTGG-CTCACGCCTGTAATCCCAGCACTTTGGGAGGCCG	49
	** **** * **** ***** ***** ****	
3'- α 1	GGTTGGGCACAGTGGCCTCACGCCTGTAATCCCAGCACTTTGGGAAGCCA	471
		(b)
promoter	GGGTCGANNCC	
Alu	AGGCGGGCGGATCACCTGAGGTCAGGAGTTC	80
	*** **** ***** ***** ***	
3'- α 1	AGGTGGGCAGATCAC--AAGGTCAGGAATTC	500

Fig. 1. (a) Sequence alignment between HY3 (2) and the corresponding 3'- α 1-globin region (1). Putative polymerase III promoter is indicated. (b) Genomic *Alu* consensus (4), aligned to the 3'- α 1-globin sequence at positions 423-500. Promoter boxes (5) are indicated. Sequences and numbering of the 3'- α 1-globin region are identical to those in (1). Exact matches (*), purine-purine/pyrimidine-pyrimidine replacements (!), and gaps (-) are indicated in both alignments.

REFERENCES

- (1) Shen, C.-K.J. and Maniatis, T. (1982) *J. Mol. Appl. Gen.* 1, 343-360.
- (2) Wolin, S.L. and Steitz, J.A. (1983) *Cell* 32, 735-744.
- (3) Hess, J., et al. (1985) *J. Mol. Biol.* 184, 7-21.
- (4) Schmid, C.W. and Shen, C.-K.J. (1985) in *Molecular Evolutionary Genetics* (McIntyre, R.J. ed.), pp. 323-358. Plenum Publishing, New York.
- (5) Fowlkes, D.M. and Shen, T. (1980) *Cell* 22, 405-413.

EVOLUTION OF HUMAN ALU REPEATS: IMPLICATIONS FOR GENOME STUDIES

J. Jurka¹ and R.J. Britten,² ¹Bionet, Mountain View, California; ²California Institute of Technology, Pasadena, California.

The human Alu family of repeated sequences contains at least five different subfamilies referred to as Alu-b,c,d,e and j which are arranged according to their position on the evolutionary time scale from the most recent, Alu-b, to the oldest, Alu-j, subfamily. This is identified by computer analysis of the correlated, subfamily-specific nucleotide occurrences in a number of the diagnostic Alu sequence positions. A specialised sequence editor has been developed to pursue these studies.

Members of each subfamily show different nucleotide preferences in the diagnostic Alu sequence positions which can easily be identified. Another characteristic feature of different subfamilies is a systematic difference in the CpG content from an average 16.08 per sequence in Alu-b to an average 3.84 per sequence in the Alu-j subfamily. It is proposed that subsequent generations of Alu subfamilies have been transcribed from different CpG-rich source genes. The genes presumably have replaced each other during the evolutionary history of primates. Once integrated into the genome, a copy of the source gene is no longer under selective pressure to maintain the original CpG content of the source gene. Therefore, one can observe a time-dependent elimination of CpG from Alu sequences.

The classification of Alu repeats, based on the diagnostic base differences and the CpG "decay", provide a method to date both the invasion of individual genes by Alu elements and the genetic rearrangements associated with this process.

Cold Spr. Harb. Symp., May 1988

1

Elzbieta Holsztyńska, David J. Waxman and Jerzy Jurka
(1) Department of Biological Chemistry and Molecular Pharmacology,
Dana-Farber Cancer Institute, Harvard Medical School; (2) Bionet,
National Computer Resource for Molecular Biology.

Currently, about 70 full-length sequences for cytochrome P-450 from 9 eukaryotic species and one prokaryote are available. The crystallographic model of bacterial cytochrome has been reported (Poulos et al., 1987, J. Mol. Biol. 195, 687) and has been used as a reference to evaluate our structural predictions using a variety of theoretical and computer methods. Postulated regions of potential structural-functional importance in P450 PB4(IIB1) include: (1) combined membrane insert halt-transfer signal residues; (2) sites of interaction with cytochrome b5 and cAMP-dependent kinase; (3) internal halt-transfer sequence; (4) large hypervariable region; (5) predicted dioxygen binding site; (6) NADPH-P450 reductase interaction site; (7) conserved region and (8) axial cysteine heme-binding region. We used Multiple Aligned Sequence Editor (Faulkner and Jurka, 1988, TIBS, in press), to elaborate a synthetic model correlating sequence variations in homologous positions of the aligned set with the functional map as well as with predicted and/or reported structural features. Results of this study will be presented.

(Protein Soc. Meeting, San Diego, August 1988)

Locating Amino Acid Patterns in Proteins by Composition

Sunil Maulik

BIONET c/o IntelliGenetics, Inc., 700 E. El Camino Real, Mountain View, CA.,
94040.

It is known that the amino acid composition of a protein plays a major role in determining its folded state [1]. A fundamental pattern-matching problem in protein analysis is that of finding sequences (or sub-sequences) given certain *compositional* criteria only. For instance, one may wish to find all regions of *unspecified length* >20% proline and >30% glycine residues in a (say) 500 amino acid peptide sequence. A related problem may be to find all hydrophobic regions (>50%) in a protein. An algorithm has been developed that will scan a sequence and find sub-sequences (of any length) satisfying given compositional criteria. The implementation of this algorithm, termed RICH, is near completion and will be available soon on BIONET. Uses of the RICH program might include finding hinge structures in immunoglobulin sequences (known to be rich in proline and cysteine residues [2]); or verifying if a protein satisfies the PEST hypothesis [3] i.e. if its half-life is related to the compositional content of P (proline) E (glutamic acid) S (serine) and T (threonine) residues.

References:

1. Sheridan RP, et al.,(1985) Biopolymers 24: 1995-2003
2. Huber R and Bennett WS (1987) Nature 326: 334-335
3. Rogers S Wells R and Rechsteiner M (1986) Science 234: 364-368

II. BIONET Training Publications

A copy of one BIONET informational publication is included in this section.

Protein databases and software on BIONET

Sunil Maulik

BIONET c/o IntelliGenetics Incorporated, 700 East El Camino Real, Mountain View, CA, 94040 USA

Abstract. BIONET provides databases, software, and networking/communications tools to over 2500 molecular biologists worldwide. Software for the analysis of nucleic acid and protein sequence data is provided by both IntelliGenetics, and academic contributors. BIONET is currently implementing dedicated high speed servers for searching protein databases, as well as providing more flexible tools for protein structure recognition and prediction. In this review, protein databases and analysis software available on the BIONET resource are described, and progress in providing new tools for structure prediction, comparative sequence analysis, and pattern recognition using Artificial Intelligence (AI) techniques are summarized.

What is BIONET?

BIONET is a national computer network for molecular biologists and biochemists. It is a non-profit resource funded by a co-operative agreement between the NIH, Division of Research Resources and IntelliGenetics, of Mountain View, California (No. 5 U41 RR01865-06). BIONET provides access to biological databases, software for analyzing the data, and communication and networking tools for the distribution of data, software, and computing resources. BIONET maintains databases and software for both nucleic acid and protein analysis. This review covers the protein analysis component of BIONET only.

An annual subscription fee of \$400 allows academic or non-profit users unlimited access to the BIONET computer (a DEC-2065 time-shared mainframe, with additional network access to a Sun 3/280 database server and a micro-VAX) including payment of telecommunication costs. Access limited only to the communications facilities is also available. International users are waived the subscription fee, but are required to pay their own telecommunication costs. Over 750 laboratories are currently subscribed to BIONET from the U.S., Europe, and Japan, corresponding to over 2500 researchers.

BIONET is directly connected to Telenet and CompuServe which provide 24-h access to the BIONET computer via a local telephone call from most cities in the United States. These networks allow scientists in Europe and Japan to access BIONET using international carriers such as Euronet, Datax-P, Transpac, and Venus. BIONET is also directly connected to the ARPAnet which provides mail and

file-transfer capabilities to any host computer on the ARPAnet and to a large number of Internet hosts on Bitnet, Usenet, CSnet, etc. This extremely high level of connectivity has facilitated both communication and collaboration between scientists worldwide. As an example, Dr. M.M. Teeter of Boston College (Teeter@bcchem.Bitnet) maintains a database of the electronic mail addresses of protein crystallographers throughout the U.S. and Europe. This database is accessible on BIONET for use with the MM mail software.

BIONET is using its extensive networking and communications facilities to maintain a number of electronic bulletin-boards (BBoards) dealing with different aspects of molecular biology. These BBoards are distributed throughout the world. Messages are exchanged with the SEQNET BBoard in Europe (SEQNET is BBoard service run by Drs. Michael Ashburner and Martin Bishop of Cambridge University), and are received as far away as Israel, Korea, Taiwan, Australia, and even the U.S. research station in Antarctica! Recently, BIONET helped initiate and participated in the formation of the worldwide BIOSCI BBoard network. Of particular interest to scientists working with proteins are the PROTEIN-ANALYSIS and METHODS-AND-REAGENTS BBoards. The former is moderated by Amos Bairoch of the University of Geneva (Bairoch@cgecmu51.Bitnet) and deals with such topics as protein chemistry, sequence analysis, and structure prediction. The latter is an all-purpose BBoard for nucleic acids and proteins, and contains requests and data on topics such as codon usage tables, peptide synthesizers, cDNA libraries, antibodies, and protein engineering/site-directed mutagenesis.

Protein databases and software

BIONET provides the Protein Identification Resource (PIR) protein sequence database [1], the SWISS-PROT protein sequence database distributed by the European Molecular Biology Laboratory (EMBL)[2], and the KeyBank™ database of protein (and nucleic acid) sequence patterns (provided by IntelliGenetics). KeyBank™ contains protein and DNA consensus sequences (motifs) described using the QUEST programs pattern language [3]. Thus, at a single resource information at several levels of biological function are immediately accessible to the researcher. IntelliGenetics provides a suite of programs that readily access information in these databases. The IFIND program will rapidly search a query sequence against any sequence databank utilizing

Ms. No. 035 Author Maulik

Ms. 1-15 Pages 1-4

Springer-Verlag, Heidelberg / H. Stürtz AG, Würzburg

Provisorische Seitenzahlen / Provisional page numbers

1. Korr.

Ⓟ

Date 23.9.88

the algorithm of Wilbur and Lipman [4], report similarity scores, and display optimized alignments between the query and similar database sequences. The QUEST program can also search any sequence databank with a given sequence pattern and locate exact matches of that pattern to sequences or subsequences in the database. QUEST may also be used in conjunction with the predefined patterns in Key-Bank[™] to locate particular subsequences (for instance, signal sequences) within a given sequence of interest.

Sequences located by either IFIND or QUEST may be simultaneously aligned by GENALIGN, a multiple sequence comparison program utilizing the regions method of Martinez [5]. GENALIGN will align up to 49 protein or nucleic acid sequences simultaneously. The program allows the user to choose between several different amino acid "alphabets" when comparing protein sequences, including the Jimenez-Montano and Zamora-Cortina alphabet of evolutionary similarity [6], the Miyata alphabet of physico-chemical similarity [7], a hydrophobic-hydrophilic alphabet, or a hydrophobic-neutral-hydrophilic alphabet. In addition, users have a flexibility to create and use their own amino acid alphabets.

IntelliGenetics' PEP program allows the user to perform various analyses on peptide sequences. These include secondary structure predictions by the Chou-Fasman algorithm [8], and hydrophobicity calculations using the Hopp-Woods [9] or Kyte-Doolittle [10] procedures. In addition, PEP can simulate chemical cleavage by a variety of proteases and chemical treatments (a database of eight common proteases and five forms of chemical cleavage exist within the program¹, and users can add or create their own database), determine amino acid composition, molecular weight, and pI. The program also performs rapid and rigorous similarity comparisons between peptide sequences. (The former using a modified [11] version of the Korn-Queen-Wegman algorithm [12], and the latter using the Needleman-Wunsch algorithm [13] as modified by Smith-Waterman [14].) PEP determines reverse translations using codon preference tables followed by restriction site mapping and splicing functions. Finally, PEP also contains a generalized "window" algorithm, that allows users to define any characteristic pattern in a sequence that can be determined using a moving weighted average (either arithmetic or geometric) over the sequence. For example, the window function in PEP may be customized to predict the existence of membrane-associated alpha-helices using the algorithm of Rao and Argos [15].

In contrast to the other programs in the IntelliGenetics suite, the SIZER program does not deal directly with sequence data, but instead may be used to calibrate gel electrophoresis bands against given standards, SIZER can use either the Duggleby [16], Southern [17], or spline [18] methods of curve fitting. SIZER may be used with either protein or nucleic acid gel fragment data.

In addition to the IntelliGenetics suite of programs for nucleic acid and protein sequence analysis, BIONET provides selected software from academic researchers. Contributed programs that deal with various aspects of protein analysis currently on BIONET include:

¹ The proteases are: trypsin, chymotrypsin, pepsin, thermolysin, clostripain, Staphylococcal protease, Myxobacter protease, and proendopeptidase. The chemical treatments are: cyanogen, bromide, hydroxylamine, iodosobenzoate, pH 2.5, and iodoacetamide

(i) FASTP, which utilizes the algorithm of Lipman and Pearson [19] to implement rapid and sensitive similarity searches of the PIR or SWISS-PROT protein databanks

(ii) The PROT3 [20], ALP3 [21] and XALIGN [22] programs for multiple sequence alignment of protein sequences

(iii) The XPROF program for the prediction of protein hydrophobicity using Rose's algorithm [23]

(iv) The IDEAS suite of structural programs [24] including DELPHI (secondary structure prediction by Robson's method [25], HPLOT (distribution of hydrophobic and charged residues using the Nozaki-Tanford [26] or Eisenberg et al. [27] methods), HCOMP (comparison of hydrophobicity profiles), and ALOM (prediction of membrane-spanning regions by discriminant analysis [28]).

(v) The DSSP program [29] for the determination of protein secondary structure from the Brookhaven Protein databank atomic coordinates.

BIONET also creates software which promotes efficient use of the resource. Recent software additions include the BIFIND and BFASTP database interactive command-file generators. BIFIND and BFASTP serve to insulate naive users from the intricacies of performing database similarity searches. They prompt the user for sufficient information to perform the database search, display menus of the different databases, and use intelligent defaults for all other parameters. They produce as their output command (batch) files, which, when submitted, instruct the appropriate database searching program (IFIND in the case of BIFIND, FASTP in the case of BFASTP) to run the searches as batch (remote) jobs.

New directions

An extremely rapid version of FASTP implemented to run on Sun workstations has been provided to BIONET by Dr. Warren Gish of U.C. Berkeley. Named FASTP-mail, it is accessible by sending an electronic mail message containing the query sequence to an IntelliGenetics Sun database-server. The program automatically reads the message and then proceeds to search the entire PIR database in as little as 30 seconds. It remains the search output (top 20 scoring database sequences as well as optimized alignments) back to the user. The implementation allows the user to proceed with other tasks on BIONET, or even log-off, while the search is taking place.

BIONET is implementing a FASTA-mail program (that searches both the GenBank/EMBL nucleotide libraries and the PIR/SWISS-PROT peptide libraries). Thanks to a generous equipment donation from Sun Microsystems, BIONET will be moving to a network of Sun workstations and file-servers during 1988. The new network will result in a greater than tenfold increase in the amount of total compute power available on BIONET. The implementation of dedicated database-servers on this network will become increasingly important as both the resource and the databases continue to expand.

BIONET has initiated a collaboration with RIACS (Research Institute for Advanced Computer Science) to use the Connection Machine II [30] for database searches. The Connection Machine II is an example of a massively parallel architecture. BIONET intends to implement software

for rapid database similarity searches and large-scale sequence alignments within this parallel environment.

In the last fifteen years protein structure prediction has become an increasingly important research topic. There is a growing demand from the BIONET community for more flexible and accurate tools for protein structure prediction and analysis. BIONET intends to build upon its expertise in comparative sequence analysis [31] and to collaborate with outside experts on protein structure recognition, analysis, and prediction to research and develop new tools in this field. Protein structure prediction by knowledge-based analysis requires as its starting point a database of structural knowledge [32] obtained from the analysis of known protein structures. Rules inferred from such a database are then combined with either pattern-matching [33] or multiple sequence alignment algorithms [34] to predict secondary structural elements, and, optimally, a single tertiary structure. If the protein sequence shares a substantial primary sequence similarity (> 50%) to a sequence of known structure, multiple sequence alignment, followed by molecular modeling, generally allows a plausible structure to be built [35, 36]. Even in the absence of known structure, predicted secondary structural elements, hydropathicity, chain flexibility, and evolutionary information combined in a multiple sequence alignment can provide increased accuracy in predicting the fold of a protein [36-38]. As obtaining optimal multiple sequence alignments algorithmically remains an unsolved problem, many researchers prefer the interactive alignment and analysis of sequences. A prominent example of software capable of such a task is the Multiple Aligned Sequence Editor (MASE) developed by Faulkner and Jurka [39], which combines interactive sequence manipulations together with standard sequence analysis functionality. MASE runs on computers under the UNIX operating system. It will be available on-line later this year when BIONET moves to the SUN network.

A complementary approach to protein structure prediction by comparative sequence analysis is one of pattern-matching. The QUEST program described earlier already provides one means of finding sophisticated patterns in sequences. However, to recognize structural features of proteins, more complex pattern searching tools are needed. A pattern language for protein structure (termed PLANS) has already been implemented [40]. An artificial-intelligence based program called MATCH has been written by R. Abarbanel. MATCH can find elements of secondary structure such as turns with better than 90% accuracy [33, 41]. Tertiary structure predictions using both pattern-matching and combinatorial approaches are now being attempted [42, 43]. BIONET intends to make MATCH and other programs developed for combinatorial prediction of protein tertiary structure available on the resource. Many of these programs result from the research of F.E. Cohen and I.D. Kuntz at UCSF [33, 40, 41] with whom BIONET intends to collaborate on disseminating research software tools. The programs as well as the knowledge-bases upon which they depend will be expanded to improve the accuracy of predictive techniques.

A related pattern-matching problem is that of finding sequences (or sub-sequences) given certain compositional criteria only. For instance, one may wish to find all regions (of unspecified length) > 20% proline and > 30% glycine residues in a 500 amino acid peptide sequence. A related problem may be to find all hydrophobic regions (> 50%)

in a protein. An algorithm has been developed that will scan a sequence and find sub-sequences (of any length) satisfying given compositional criteria. The implementation, termed RICH, is in progress. (Details of the algorithm and its implementation will be described elsewhere.) Uses of the RICH program might include finding hinge regions in immunoglobulin sequences (known to be rich in proline and cysteine residues [44]); or verifying if a protein satisfies the PEST hypothesis [45], i.e., if its half-life is related to the compositional content of P (proline) E (glutamic acid) S (serine) and T (threonine) residues.

Summary

With efforts underway to sequence complete genomes, the number of actual and deduced protein sequences will increase rapidly. Sequence databases and software must become integrated with "higher order" databases (e.g., of protein structural characteristics). Suitable software must also be developed to link them together. BIONET already acts as a central repository for sequence databases, sequence and pattern searching software, and electronic media for the acquisition and dissemination of new biological information. Additionally, BIONET is actively pursuing research to create new algorithms to allow the functional characteristics of protein molecules to be deduced from their sequences.

References

- George DG, Baker WC, Hunt LT (1986) NAR 14:11-16
- Hamm GH, Cameron GN (1986) NAR 14:5-10
- Abarbanel RA (1984) NAT 12:263-280
- Wilbur WJ, Lipman DJ (1983) Proc Natl Acad Sci USA 80:726-730
- Sobel F, Martinez HM (1986) NAR 14:363-374
- Jimenez-Montano M, Zamora-Cortina L (1981) Proceedings, VII International Biophysics Congress, Mexico City
- Miyata T, Miyazawa S, Yasunaga T (1979) J Mol Evol 12:219-236
- Chou PY, Fasman GD (1974) Biochemistry 13:211-245
- Hopp TP, Woods KR (1981) Proc Natl Acad Sci USA (1981) 78:3824-3828
- Kyte J, Doolittle RF (1982) J Mol Biol 157:105-119
- Brutlag D, Clayton J, Friedland P, Kedes LH (1982) NAR 10:279-294
- Korn LJ, Queen CL, Wegman MN (1977) Proc Natl Acad Sci USA 74:4401-4405
- Needleman SB, Wunsch CD (1970) J Mol Biol 48:443-453
- Smith TF, Waterman MS (1981) J Mol Biol 147:195-197
- Rao MJK, Argos P (1986) Biochem Biophys Acta 869:197-214
- Duggelby R, Kinns H, Rood JI (1981) Anal Biochem 110:49-55
- Southern EM (1979) Anal Biochem 100:319-323
- Vandergraft JS (1983) Spline interpolation. In: [] [] (eds) Introduction to numerical computations, 2nd ed. Academic Press, San Francisco, pp 126-138
- Lipman D, Pearson W (1985) Science 227:1435-1441
- Murata M, [] [] [] (1985) PNAS 82:3073-3077
- Gotoh O (1986) J Theor Biol 121:327-337
- Bacon DJ, Anderson WJ (1986) J Mol Biol 191:153-161
- Rose GD, [] [] [] Science (1985) 229:834-838
- Kanehisa M (1984) IDEAS. Integrated database and extended analysis system for nucleic acids and proteins. User manual. Laboratory of Mathematical Biology, NIH, Bethesda, MD
- Garnier J, Osguthorpe DJ, Robson B (1978) J Mol Biol 120:97-120
- Nozaki Y, Tanford C (1971) J Biol Chem 246:2211-2217

14.

- 4
- 15
27. Eisenberg D, Weiss RM, Terwilliger TC (1984) *Proc Natl Acad Sci USA* 81:140-144
 28. Klein P, Kanehisa M, DeLisi C (1985) *Biochem Biophys Acta* 815:468-476
 29. Kabsch W, Sander C (1983) *Biopolymers* 22:2577-2637
 30. Hill D (1985) *The connection machine*. MIT Press, Cambridge, MA
 31. Jurka J, Smith TF (1988) *Proc Natl Acad Sci USA* 85:4775-4778
 32. Blundell TL (1987) *Nature* 326:347-352
 33. Cohen FE, ■■■ ■■■ ■■■ (1983) *Biochemistry* 22:4894-4904
 34. Webster TA, Lathrop RH, Smith TF (1987) *Biochemistry* 26:6950-6957
 35. Lesk AM, Chothia CH (1986) *Phil Trans R Soc London Ser A* 317:345
 36. Brown JH, ■■■ ■■■ ■■■ (1988) *Nature* 332:845-850
 37. Crawford IP, Niermann T, Kirschner K (1987) *Proteins* 118-129
 38. Pearl LH, Taylor WR (1987) *Nature* 329:351-354
 39. Faulkner DV, Jurka J (1988) *Trends Biochem Sci* ■■:■■-■■
 40. Abarbanel RA (1984) Ph.D. thesis, University of California, San Francisco
 41. Cohen FE, ■■■ ■■■ ■■■ (1986) *Biochemistry* 25:266-271
 42. Cohen FE, ■■■ ■■■ ■■■ (1986) *Science* 234:349-352
 43. Webster TA, Lathrop RH, Smith TF (1987) *Biochemistry* 26:6950-6957
 44. Huber R, Bennett WS (1987) *Nature* 326:334-335
 45. Rogers S, Wells R, Rechsteiner M (1986) *Science* 234:364-368

Received May 31, 1988 / Accepted June 27, 1988

III. BIONET Newsletters

Copies of the two BIONET Newsletters are included in this section.

BIONET NEWS

vol. 1 no. 1

April 1988

Automatic Data Submission to GenBank, EMBL, and NBRF-PIR

BIONET users can conveniently submit sequence data using the XGENPUB program. XGENPUB helps to annotate and electronically submits sequence data to GenBank, the National Institutes of Health DNA sequence library, EMBL, the nucleotide sequence database from the European Molecular Biology Laboratory, and NBRF-PIR, the National Biomedical Research Foundation's protein sequence database.

Sequence authors are encouraged to submit their sequence data once it has been derived. Direct computer-readable author submissions decrease the time delay with which the data is incorporated into the databases and can improve the completeness and accuracy of the annotation and sequence data.

XGENPUB is located on-line and can be accessed by simply typing "XGENPUB" at the system @ prompt. The program accepts sequence data in files from the author's BIONET directory and prompts the user for the name of the sequence file and the sequence name. XGENPUB initiates an editor for completing a submission form and inserts the sequence data at the end of the form. When the user exits from the session, the completed entry is automatically mailed to GenBank, EMBL, and NBRF-PIR using the ARPANET computer network. A copy of the submission will also be sent to the author's mail file. Sequence data will be present in the databases within four months of submission.

The vast rate of growth that databases are experiencing moves sequence submission responsibilities to the authors. Many journals are now requesting that their contributing authors submit the data. One journal, *Nucleic Acids Research*, requires evidence that data have been submitted before it will consider a manuscript for review. The XGENPUB program provides BIONET users with a convenient method with which to fulfill their sequence

submission responsibilities. Further information regarding XGENPUB may be obtained by viewing the on-line help topic HELP XGENPUB. □

Efficient Database Searching

In an effort to increase efficient utilization of the resource, BIONET has produced "Command File Generators." These programs create a formatted list of program commands for remote operation of the sequence database searching software on BIONET. BIFIND, BFASTP, and BFASTN serve to insulate users from the intricacies of the IFIND, FASTP and FASTN database similarity searching programs. Each prompts the user for sufficient information to perform the similarity search, displays menus of the different database choices, and uses intelligent defaults for all other parameters. They then produce as their output command (batch) files, which, when submitted, instruct the appropriate database searching program to run the search as a batch or remote job. The user may then safely log-off the BIONET system, knowing that the output of the database search will be stored in a file for subsequent analysis once the batch job has run. On-line help exists for all the programs by typing HELP followed by the program name.

The Command File Generators undergo continued development. As an example, BIFIND will soon be able to save BOTH search and alignment results. With the advent of BIONET's new FASTP-mail database server (that searches the entire PIR protein databank in as little as 30 seconds - 3-5 times faster than any other implementation of FASTP), a new BFASTP is being implemented, that makes use of this new server software. Watch also for a FASTN-mail program capable of searching the GenBank nucleic acid sequence database far more rapidly than previously capable, and a corresponding new BFASTN program. □

Revised Introduction to BIONET

In May, a revised version of the *Introduction to BIONET* will be sent to all of the Principal Investigators subscribing to

BIONET. The *Introduction to BIONET* is the handbook sent to new subscribers which provides basic information for accessing and using BIONET.

The new version includes two new sections which should benefit users greatly. A "Troubleshooting" chapter helps users diagnose and solve commonly encountered technical problems. A "Program Examples" appendix provides annotated examples of thirteen common biological tasks and how the BIONET programs may be used to solve them. Many more program examples will be available in printed form when the new IntelliGenetics User Manual, described in a separate article, is distributed to BIONET users this spring.

Revisions to the *Introduction to BIONET* include the addition of a step-by-step example of a BIONET user's first login session. It covers how to connect, log into the BIONET account, read login messages, and get help. The chapter describing the BIONET programs now includes a complete list of the contributed molecular biology software that is available on-line or through the BIONET Lending Library of diskettes. The chapter on databases has been expanded to reflect the new organization of the databases as well as recently added databases such as SWISS-PROT. This chapter also describes the new XGENPUB program which allows electronic submission of sequence data and annotations to all of the principal databases. Much effort has been given to incorporating user suggestions for changes into the new Intro. We hope that it allows users to spend less time learning, and more time using the BIONET resource. □

BIONET News is published by the BIONET Computer Resource staff. Contributors for this issue were Kathryn Berg, Vickie Johncox, David Kristofferson, Sunil Maulik and Spencer Yeh. Your comments and suggestions on this publication would be greatly appreciated.

BIONET Communications Reach Worldwide

The BIONET National Computer Resource for Molecular Biology is far more than merely a facility to analyze DNA sequences. In addition to its analysis software and molecular biology databases, the Resource offers access to powerful electronic communications networks with worldwide scope. Use of this facility is completely free to BIONET account holders.

BIONET maintains the most extensive network available of molecular biology electronic bulletin boards (bboards). Nineteen bulletin boards are maintained on the BIONET mainframe computer and copies of posted messages are read around the world. The BIONET bboards provide scientists with the means of contacting a large community of colleagues by simply sending a single electronic mail message. A detailed list of bboards can be seen on BIONET by typing HELP BB-LIST after the @ prompt.

BIONET bulletins are distributed via the ARPANET, BITNET, and USENET computer networks and can be received by scientists without BIONET accounts who have electronic mail addresses on any of these networks. There is no charge for this service. Users on any of these networks can also post messages directly to any of these bboards without editorial intervention. One simply uses the address format BBOARDNAME@BIONET-20.ARPA. BIONET users can post messages to the bboards by simply entering the bboardname in response to the To: prompt when sending a message in the MM mail program.

Copies of each message are automatically relayed around the world via our direct network contacts and through our message exchange agreement with the SEQNET bboard service in Cambridge, England. Messages are read by scientists in places as far away as Japan, Korea, Taiwan, Israel, and Australia.

If there are many people working on a local campus computer at your institution who may be interested in participating in the bboards, please have the person in charge of your computer facility contact kristofferson@bionet-20.arpa. We can then arrange an efficient redistribution scheme at your site.

Electronic bulletin boards are just beginning to be discovered by the molecular biology

community. Bboards truly have the potential to revolutionize the way that science is done in this field! □

Addressing Electronic Mail Outside of BIONET

Many BIONET users have colleagues with BITNET or EARN (European Academic Research Network) electronic mail addresses and wonder how to send mail to them. The procedure is extremely simple, and there are no charges for sending mail messages. One sends a message using the normal BIONET mail routine; only the address for the message is slightly different. Suppose that a colleague's BITNET address is SMITH@YALEVM. Enter SMITH@YALEVM.BITNET in response to the To: prompt in the MM mail program:

```
@mm
MM> send
To: smith@yalevm.bitnet
```

The procedure is the same for EARN addresses. SMITH@EMBL becomes SMITH@EMBL.EARN. Actually SMITH@EMBL.BITNET will also work for EARN addresses, since the two networks are essentially the same. EARN is BITNET in continental Europe.

Some users have trouble addressing mail to English addresses. The network used in England is called JANET (Joint Academic Network). Unfortunately, JANET addresses are arranged in reverse order of that used in many other countries, including the US. To circumvent this problem, BIONET has implemented the JANET address extension (properly termed a "pseudo-domain"). To send to a JANET address mad3@uk.ac.cam.bio, add JANET to the end of the address:

```
mad3@uk.ac.cam.bio.janet
```

Note that case does not matter in mail addresses.

Finally, sometimes one encounters troublesome addresses that contain !'s: foop!sop!net@pri.com. When entering these addresses at the To: prompt, enclose the part of the address to the left of the @ in quotes. Enter:

```
"foop!sop!net"@pri.com
```

All of the !'s should be enclosed between the two quotation marks for the address to be properly interpreted by the BIONET mail software.

This short article should provide the

means to contact users almost anywhere in the networked world provided one has their electronic address. Of course, as with regular mail, one somehow has to find that address by other means. International user directories are still in their infancy, but there are means of determining addresses if the telephone or regular mail can not be used to find this. The BIONET consultants can help with this and any other communications questions. Please do not hesitate to send them an e-mail message to the BIONET address or call them at (415)324-4363. □

User Manual Replaces Short Course

A User Manual, which replaces the current Short Course, will be made available to BIONET Principal Investigators this Spring. Each Principal Investigator will receive one copy of the manual.

The User Manual is organized into the following sequence analysis topics:

- sequence entry and editing
- sequence location
- sequence and map display
- sequencing project management
- sequence translation
- sequence composition
- sequence structure
- restriction analysis
- cloning simulation
- sequence comparison and alignment

Each topic contains annotated program examples illustrating the step-by-step procedures involved in performing specific operations. The User Manual also includes sections on file structure, databases, and program commands. Additional information may be obtained from Kathryn Berg at (415)962-7337. □

Subscription Fees

Each year BIONET account holders are assessed a flat \$400.00 fee used to cover telecommunications charges. New accounts are billed when the BIONET account is activated. For renewing subscribers, a reminder notice is sent out, followed by an invoice a few weeks later. It would be appreciated if the invoices for both new accounts as well as renewing subscribers would be paid within a reasonable length of time.

We have instituted a policy of freezing delinquent accounts if the yearly subscription fee is not forthcoming. If your account has been frozen, please call (415)962-7337. □

BIONET NEWS

vol. 1 no. 2

October 1988

BIONET Grant Renewal

The grant for the BIONET National Computer Resource for Molecular Biology is currently up for renewal. We have been advised by the NIH that it is appropriate for us to solicit testimonial letters about the Resource from our user community. While we have collected many comments sent in by users over the last five years, formal testimonial letters are preferred. We request that you take the time as soon as possible to send us your written comments about BIONET (signed hardcopy, please!). Please detail both the positive and negative aspects of the Resource to enable us to assess the service that we provide. Our mailing address is:

Dr. David Kristofferson
BIONET
c/o IntelliGenetics
700 E. El Camino Real
Mountain View, CA 94040

We would also appreciate three copies of any publications from your laboratory since December 1, 1987 that have entailed the use of BIONET. Thank you in advance for your cooperation.

New FASTA-MAIL Program Speeds Similarity Searches

As the first step towards utilizing the new BIONET Sun computers, the FASTA-MAIL program, developed by BIONET's systems programmers (Eliot Lear and Rob Liebschutz), was released August 17. Since then, over 800 GenBank, PIR, and SWISS-PROT searches have been run by BIONET users. The FASTA-MAIL program allows a user to send a mail message containing a nucleic acid or protein query sequence to a BIONET Sun computer. A sequence similarity search is then performed against a nucleic acid or protein databank using the FASTA program developed by William Pearson

and David Lipman. The results are sent to the user's mail file on the BIONET DEC-2065 computer. The FASTA program is an improved version of the older FASTP and FASTN programs. By incorporating distinct scoring matrices, both protein and DNA searches are executed by the same program. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. Thus gap-containing sequences score higher on the first pass, and are retained for further consideration and optimized alignment.

To use FASTA-MAIL on BIONET, the query sequence must adhere to standard IntelliGenetics file format conventions. Because FASTA-MAIL is a mail server, two other stipulations also apply: (1) no line in the file can be longer than 80 characters and (2) there can only be one sequence in the file. To start the FASTA-MAIL program, type FASTA-MAIL at the BIONET "@" prompt. The program is quite simple and user-friendly, but we do urge users to read the on-line HELP FASTA-MAIL help topic at the "@" prompt. This documentation explains the FASTA scoring procedure, gives practical advice about submitting jobs, and contains bibliographic references for the program. A second help topic, HELP EX-35, provides a complete step-by-step example of the program from submission of a FASTA-MAIL job through a review of the results. Parameters that the user chooses are which database to search and the value of "KTUP". The KTUP value is analogous to the "WORDSIZE" parameter in the IFIND program. It determines how many consecutive residues must match in the first pass before the region is considered by the program. A lower KTUP value increases the sensitivity of the search, but lengthens the search time exponentially. The turnaround time on searches depends primarily on how many other jobs are in the queue and the KTUP value. Currently, jobs seem to be

running in 10 to 20 minutes for protein searches and in 1 to 3 hours for full GenBank searches.

Many users have asked questions regarding the interpretation of search results. Bear in mind that a high similarity score cannot prove biological homology, it can only provide evidence. FASTA-MAIL provides three scores for each sequence displayed, INITN, INIT1, and OPT. INIT1 is the same as the old FASTP or FASTN "INIT" score and is primarily included in FASTA-MAIL for comparison with results from these older programs. The INITN score is the score used on the first pass to rank sequences or discard those below a cutoff score. While the INITN score does allow gaps within matching regions, the gap-penalty does not vary with distance at this point, nor does the program check to see if there is a better alignment of the two sequences yielding a better score. These are included when the final OPT score is calculated by a Needleman-Wunsch/Smith-Watterman-type of alignment on the region surrounding the highest scoring initial region. The alignment displayed is based on this final alignment procedure. Thus, while the INITN score determines which sequences are kept and the order in which they are displayed, the OPT score gives a more precise gauge of the similarity of sequences. However, since the OPT score is not calculated for each sequence in the database, there is no mean and standard deviation generated for the OPT scores. The INITN score still must be used when one compares the high scoring sequences against the full population.

Two methods are useful for determining statistical significance. One can simply compare the INITN score of the sequence in question with the mean and standard deviation of all INITN scores to see how many standard deviations above the mean the score lies. However, one must keep in mind that in a sufficiently large database search there usually will be some random matches that still score even three or four standard deviations above

the mean. To see if a high scoring sequence is truly similar to the target query, Lipman and Pearson suggest randomizing one of the two sequences and then scoring it with the same INITN or OPT scoring procedure. If the original unrandomized score is much higher than randomized runs, this indicates that the high score is related to the sequence order and not just to the sequence composition. William Pearson has developed a program, RDF2, which performs this Monte Carlo simulation. This program will be available on BIONET when users are shifted to the Sun system. Contact Dr. Pearson directly if you wish to obtain RDF2 for PC's or mainframe Unix computers.

The original paper by William Pearson and David Lipman describing the FASTA program, "Improved tools for biological sequence comparison," appeared in the April 1988 issue of PNAS (Vol. 85, pp. 2444-2448). William Pearson may be contacted by e-mail at "wrp@virginia.edu" or at the Department of Biochemistry; Box 440, Jordan Hall; Univ. of Virginia; Charlottesville, VA 22908. BIONET would like to thank William Pearson and David Lipman for allowing BIONET users access to their programs and for providing details regarding the program's scores. We look forward to receiving comments and suggestions about the FASTA-MAIL program.

Multiple Aligned Sequence Editor (MASE)

Analysis of large sequences or large number of sequences can be quite tedious, frustrating and, above all, time-consuming. Most algorithms can only handle sequences of limited size. Outputs from available multiple sequence alignment algorithms often need further refinements to make biological sense. These and related problems with sequence analysis stimulated my joint effort with Don Faulkner to design and develop a general-purpose multiple sequence editor now called MASE. Our work began back in 1986 when I was a research fellow at Harvard. It has been continued as a collaborative project since I joined BIONET as a scientist in the middle of 1987. An introductory article on

MASE can be found in the August issue of *Trends in Biochemical Sciences*, (Faulkner, D.V. & Jurka, J.; vol. 13, pp. 321-322). A limited number of reprints of this article are still available from either co-author.

MASE can be installed on computers running the Berkeley UNIX operating system. Many non-profit research groups with access to a Berkeley Unix system have already acquired the editor free of charge from Harvard. For those non-profit researchers who do not have access to the appropriate hardware, BIONET is authorized to distribute MASE on-line. Currently, BIONET is moving from our DEC-20 to a network of donated SUN computers. After this transition, MASE will be made available on-line in late 1988 or early 1989. From the technical point of view it would be very useful to know how many potential BIONET subscribers need the sequence editor. We therefore ask potential new users to mail an electronic message to jurka@bionet-20.bio.net. Thank you for your help.

Jerzy Jurka

November Training

Are you a new user of BIONET? Or are you an experienced user wanting to become more proficient with your use of BIONET? Plan to participate in an upcoming BIONET training session.

The next class is scheduled for November 17-18, 1988. Topics covered will include sequence data entry and editing, sequencing gel management, nucleic acid and peptide sequence analysis, database structure and sequence retrieval, collecting sequences and searching for patterns, sequence similarity searches, and electronic mail.

The training session will run from 8:30 am until 5:00 pm and from 6:30 until 9:00 pm on November 17th. The evening session will cover the TOPS20 operating system, communications software, the on-line help facilities, as well as an overview of contributed software. The session on November 18th begins at 8:00 am with a question and answer section. Formal instruction begins at 9:00 am and concludes at 5:00. Lunch is provided each day. Evening meals are not

provided, but there are reasonably priced restaurants within walking distance.

BIONET is located in the IntelliGenetics Building, 700 East El Camino Real (at the intersection of Highway 85) in Mountain View, CA. This is within easy driving distance from either the San Francisco or San Jose airports.

The cost of the training is kept low to encourage users to attend. The regular price is \$100 for the two day program. To help offset the additional cost of transportation, out-of-state subscribers pay \$50.00.

The training is limited to 12 people on a first-come, first-serve basis. To reserve a place, call Kathryn Berg at (415) 962-7337. We accept purchase orders, MasterCard, VISA, or personal check. Payment should be sent to:

BIONET
c/o IntelliGenetics
700 East El Camino Real
Mountain View, CA 94040

We are enthusiastic about the BIONET resource and look forward to facilitating your use of it.

1989 Training Schedule

Listed below are the dates for BIONET training in 1989. Note that the course has been expanded to cover additional topics, such as contributed software, file transfers and telecommunications.

February 6, 7*
March 15, 16, 17
May 17, 18, 19
July 19, 20, 21
September 20, 21, 22
November 15, 16, 17
*evening session on 2/7

BIONET News is published by the BIONET Computer Resource staff. Contributors to this issue were Kathryn Berg, Karen Davis, Jerzy Jurka, David Kristofferson, and Spencer Yeh. Your comments and suggestions on this publication are greatly appreciated.

IV. BIONET Software Lending Library Catalog

A copy of the catalog is provided in this appendix.

PC-Lending Library Software

Please send one disk for each program requested. Include a self-addressed, stamped mailer along with the disks and request form. This will greatly streamline the ordering procedure. Send your request to BIONET, 700 E. El Camino Real suite 300, Mountain View, CA 94040. Substantive questions about the programs, or any difficulties with program functions should be addressed to bionet@BIONET-20.arpa.

I. Communications/Terminal Emulation

Multipurpose public-domain software for logging on to BIONET, VT100 emulation and file transfer.

Various formats are available:

- Kermit--for IBM PC compatible computers
- MacIIKermit and Kermit Shareware--for Apple Macintosh, Macintosh II, and Macintosh SE computers (send 2 disks for these programs)

II. Editor

- Micro-emacs--text editor available for BIONET subscribers. Either high or low density versions may be requested.

III. Contributed Software

All programs come on 5 1/4 " diskettes

SEQAIDII

D. J. Roufa and D. D. Rhoads
Division of Biology
Kansas State University
Manhattan, KS 66506

Droufa@BIONET-20.arpa

SEQAIDII is a multifunctional program for DNA sequence analysis.

System requirements: IBM PC compatible computer. Working files require 200 000 bytes of disk storage space.

Files Included:

- INSTALL.DOC--Documentation file
- SEQAIDFD.EXE--Self-extracting archive of SEQAIDII files
- SEQAIDII.NEW--Release notes for version 3.0
- README.DOC--Installation instructions

PCZUCKER

Michael Zuker
 National Research Council
 Biological Sciences
 Room 3115
 100 Sussex Dr.
 Ottawa, ON K1A 0R6 CANADA

Zuker@BIONET-20.arpa

PCZucker is a program for global prediction of RNA secondary structure.

System requirements: IBM PC compatible computer

The programs require a minimum of 512 kbytes of RAM, version 2.0 (or later) of DOS, and a floppy containing 360 kbytes, or 1.2 mbytes.

Files Included:

- Two versions of PCFOLD
 1. PCFOLD.EXE--packs the arrays, to extend the length of the longest fragment it can fold (425 bases maximum).
 2. PCFOLD2.EXE--does not contain packing, can handle only fragments of 345 bases maximum.
- README.DOC--Documentation and explanation of program use
- MENU.DAT,MENU.DAT2--Data files needed by the program
- FOLD.ENR--Energy file
- FOLD.BAT--Batch file to run PCFold program
- FOLD2.BAT--Batch file to run PCFold2 program
- PSTV--Default sequence file

The source code is available in:

- SOURCES.1--Source programs for PCFOLD.EXE
- SOURCES.2--Source programs for PCFOLD2.EXE

MOLECULE

John Thompson
 Carnegie-Mellon University
 Biological Sciences
 616 Mellon Institute
 4400 Fifth Ave
 Pittsburgh, PA 15213

Woolford.Thompson@BIONET-20.arpa

This disk contains compressed files for John Thompson's MOLECULE program for display of secondary structure prediction. The programs read .CT files produced by Zuker's PCFOLD program, and display a 2D graphic representation of the structure. Three versions of MOLECULE are available, but only one need be accessed, depending on the monitor-type being used.

System requirements: IBM PC compatible computer

Files Included:

- CMOLECULE.ARC--IBM colorgraphics monitor
- EMOLECULE.ARC--EGA monitor (Enhanced Graphics Adaptor)
- HMOLECULE.ARC--Hercules Monochrome Card
- *.DOC--instructions on program use
- *.PAS--source code, in Turbo Pascal

ALIGN

Dan Davison
 Dept of Biochemical and Biophysical Sciences
 University of Houston
 University Park
 4800 Calhoun
 Houston, TX 77004

Goad.Davison@BIONET-20.arpa

Keith Thompson
 Biology Dept.
 Brookhaven Nat'l Lab
 Long Island, NY

The ALIGN.DOC file provides a detailed description of how the program compares sequences the user has submitted. The documentation includes a step-by-step explanation of the procedures involved, and clarifies program parameters. Several types of output are available. The program has the capacity to print any or all of the following:

1. input data (amino acid or nucleotide sequences typed in by user)
2. table of matches--which show actual start and stop positions of each match found
3. a listing of matched and unmatched areas

System requirements: IBM PC compatible computer

Files Included:

- ALIGN.EXE--executable version of ALIGN
- ALIGN.DOC--documentation and explanation of program use

There are many other accessory files also on this disk.

ALP3/ALN3

Osamu Gotoh
 Dept. of Biochemistry
 Saitama Cancer Center Research Institute
 Ina-machi Saitama 362
 JAPAN

This diskette contains some implementations of the algorithm for aligning three protein or DNA sequences described in Gotoh, O. (1986) J. Theor. Biol. 121, 327-337.

System requirements: IBM PC compatible computer

Files included:

- ALN3.EXE--program for aligning three DNA sequences
- ALP3.EXE--program for aligning three protein sequences
- MDM-1.DAT--program data file
- MAKMDM.EXE--accessory file

Test data files:

- S1.SEQ DNA
- S2.SEQ "
- S3.SEQ "
- P1.PEP Protein
- P2.PEP "
- P3.PEP "

Also included:

- SEQFORM.DOC --describes the sequence file format to use when running ALN3/ALP3.

PLASMID PAINT

Joe Lipsick
 Dept. of Pathology, M-012
 U C San Diego
 La Jolla, CA 92013

Jlipsick@BIONET-20.arpa

PLASMID PAINT, a program written in Microsoft QuickBASIC 2.0, allows one to draw plasmids.

System requirements: IBM PC compatible computer with a CGA-compatible adapter and screen

Files included:

- PLASMIDC.EXE--executable file
- PAINT.BAT--short batch file which loads the DOS GRAPHICS command and then runs PLASMIDC.EXE.
- README.DOC--explanation of program use

OLIGO MUTANT MAKER

Kevin Beadles
1044 1/2 Shrader St.
San Francisco, CA 94117

(415)759-0148 (Kevin will call you back collect if he must return your call.

OligoMutantMaker simplifies the designing and screening of oligonucleotide-directed single amino acid substitution experiments by searching for nucleotide sequences which introduce a restriction endonuclease recognition sequence into the codon substitution site of the mutant.

System requirements: IBM PC compatible computer.

Files Included:

- README.DOC--Documentation file
- BWMUTANT.COM--Executable file for monochrome monitors.
- CMUTANT.COM--executable file for color monitors.
- CUTTERS.DAT--Binary database of enzyme cleavage and availability data.
- CODONID.DAT--Standard genetic code.
- [ENZYME.TXT]--This file is created every time the program is run. It contains a copy of the results of the last analysis.

V. BIOSCI Bulletin Board Network Information

A copy of the information sheet which is mailed electronically to interested parties is included in this section.

BIOSCI BULLETINS

The following is a list of bboards available for distribution to sites on the ARPANET/Internet, BITNET, EARN, NETNORTH, and JANET. For each of these bboards a list of BITNET name abbreviations and analogous USENET newsgroup names are listed below. Finally, we provide a list of the various sites (nodes) that distribute the bboards and the address format for posting messages. Note that messages posted at any node are automatically redistributed to all other nodes on the BIOSCI network and subsequently to their readers.

BBOARD NAME	TOPIC
-----	-----
AGEING	Scientific Interest Group
BIONEWS	General announcements
BIOTECH	Biotechnology issues
BIO-CONVERSION	Scientific Interest Group
BIO-MATRIX	Applications of computers to biological databases
CONTRIBUTED-SOFTWARE	Information on molecular biology programs contributed to the public domain
EMBL-DATABANK	Messages to and from the EMBL database staff
EMPLOYMENT	Job opportunities
GENBANK-BB	Messages to and from the GenBank database staff
GENE-EXPRESSION	Scientific Interest Group
GENOMIC-ORGANIZATION	Scientific Interest Group
METHODS-AND-REAGENTS	Requests for information and lab reagents
MOLECULAR-EVOLUTION	Scientific Interest Group
ONCOGENES	Scientific Interest Group
PC-COMMUNICATIONS	Information on communications software
PC-SOFTWARE	Information on PC-software for scientists
PIR	Messages to and from the PIR database staff
PLANT-MOLECULAR-BIOLOGY	Scientific Interest Group
PROTEIN-ANALYSIS	Scientific Interest Group
RESEARCH-NEWS	Research news of interest to the community
SCIENCE-RESOURCES	Information about funding agencies, etc.
SWISS-PROT	Messages to and from the SWISS-PROT database staff
YEAST-GENETICS	Scientific Interest Group

BITNET abbreviations (<= 8 characters) for each bboard have been established:

BBOARD NAME	BITNET/EARN Name
-----	-----
AGEING	AGEING
BIONEWS	BIONEWS
BIOTECH	BIOTECH
BIO-CONVERSION	BIO-CONV
BIO-MATRIX	BIOMATRX
CONTRIBUTED-SOFTWARE	SOFT-CON
EMBL-DATABANK	EMBL-DB
EMPLOYMENT	BIOJOBS
GENBANK-BB	GENBANKB
GENE-EXPRESSION	GENE-EXP
GENOMIC-ORGANIZATION	GENE-ORG
METHODS-AND-REAGENTS	METHODS
MOLECULAR-EVOLUTION	MOL-EVOL
ONCOGENES	ONCOGENE
PC-COMMUNICATIONS	SOFT-COM
PC-SOFTWARE	SOFT-PC
PIR	PIR-BB
PLANT-MOLECULAR-BIOLOGY	PLANT

PROTEIN-ANALYSIS
 RESEARCH-NEWS
 SCIENCE-RESOURCES
 SWISS-PROT
 YEAST-GENETICS

PROTEINS
 RESEARCH
 SCI-RES
 SWISSPRT
 YEAST

Note: For the database bboards addresses such as PIR, EMBL, etc., were avoided since these may conflict with other BITNET addresses at the EMBL, etc., if they join the distribution scheme.

Equivalences of the Unix USENET newsgroup names to the ARPANET mailing list names follow:

```

bionet.general .....BIONEWS
bionet.jobs .....EMPLOYMENT
bionet.technology.general .....BIOTECH
bionet.technology.conversion .....BIO-CONVERSION
bionet.molbio.news .....RESEARCH-NEWS
bionet.molbio.ageing .....AGEING
bionet.molbio.bio-matrix .....BIO-MATRIX
bionet.molbio.methds-reagnts .....METHODS-AND-REAGANTS
bionet.molbio.genbank .....GENBANK-BB
bionet.molbio.embl databank .....EMBL-DATABANK
bionet.molbio.pir .....PIR
bionet.molbio.evolution .....MOLECULAR-EVOLUTION
bionet.molbio.gene-express .....GENE-EXPRESSION
bionet.molbio.gene-org .....GENOMIC-ORGANIZATION
bionet.molbio.oncogenes .....ONCOGENES
bionet.molbio.plant .....PLANT-GENETICS
bionet.molbio.proteins .....PROTEIN-ANALYSIS
bionet.molbio.swiss-prot .....SWISS-PROT
bionet.molbio.yeast .....YEAST-GENETICS
bionet.sci-resources .....SCIENCE-RESOURCES
bionet.software.pc .....PC-SOFTWARE
bionet.software.pc.comm .....PC-COMMUNICATION
bionet.software.contrib .....CONTRIBUTED-SOFTWARE
  
```

BIOSCI Nodes

Information about the BIOSCI bboard network can be obtained by mailing to the address

biosci@xxxx

where xxxx can be any of the following node addresses, e.g., biosci@uk.ac.daresbury in the United Kingdom. Interested parties outside of Europe and North America should contact whichever node is most convenient. Messages can be posted directly to bboards at any of these nodes by using the address format

bboard@xxxx

where bboard is a name from the lists above and xxxx is from the node list below, e.g., bionews@umdc.bitnet Note that nodes listed as (Internet) sites utilize the long bboard names as indicated in the first list above and nodes listed as BITNET, EARN, or JANET use the BITNET abbreviated bboard names.

Europe

```

-----
Sweden    bmc.uu.se      (Internet)
UK         uk.ac.daresbury (JANET)
Ireland   irlearn.bitnet (BITNET/EARN)
Ireland   irlearn.ucd.ie (Internet,
              but uses BITNET bboard names)
  
```

North America

```

-----
net.bio.net (Internet)
bionet-20.bio.net (Internet)
umdc.bitnet (coming soon, BITNET)
  
```

VI. BIONET Training Schedules

The schedules for the five Mountain View Trainings are included.

Training Schedule

March 17 & 18, 1988

Thursday, March 17

Time	Topic	Instructor
9:00- 9:45	Introduction	Nancy Bigham
9:45-10:15	Sequence Entry	
10:15-10:30	Break	
10:30-12:00	Sequencing Project Management	Vicki Johncox
12:00- 1:00	Lunch	
1:00- 1:20	Cloner Demonstration	Constance Gertsch
1:30- 2:00	Sequence Alignment	David Kristofferson
2:00- 3:15	DNA Sequence Analysis	
3:15- 3:30	Break	
3:30- 4:45	Peptide Sequence Analysis	Spencer Yeh

Friday, March 18

Time	Topic	Instructor
9:00- 10:30	Database Structure and Simple Searches	Constance Gertsch
10:30-10:45	Break	
10:30-12:00	Finding Sequences in the Databases	
12:00- 1:00	Lunch	
1:00- 3:00	Sequence Alignment	Sunil Maulik
3:00- 3:15	Break	
3:15- 4:00	Electronic mail and bulletin boards	Dave Kristofferson

**An Introduction to the IntelliGenetics Suite
Course Schedule
May 19 and 20, 1988**

Thursday, May 19

8:30	9:00	Introduction Overview of the IntelliGenetics Programs	Nancy Bigham
9:00	9:45	Sequence Data Entry and Editing	
9:45	10:30	System commands	
10:30	10:45	Break	
10:45	12:00	Sequencing Gel Management	Vicki Johncox
12:00	1:00	Lunch	
1:00	2:15	Nucleic Acid Sequence Analysis	Trish Benton
2:15	2:30	Break	
2:30	3:45	Peptide Sequence Analysis	Spencer Yeh
3:45	5:00	Electronic Mail and Bulletin Boards File Transfer	Dave Kristofferson

Friday, May 20

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Beth Swank
10:30	10:45	Break	
10:45	12:00	Collecting related sequences, searching for patterns	
12:00	1:00	Lunch	
1:00	3:15	Sequence Similarity Searches	Sunil Maulik
3:15	3:30	Break	
3:30	4:30	Review and Questions	Nancy Bigham

**An Introduction to the IntelliGenetics Suite
Course Schedule
July 14 and 14 1988**

Thursday, July 19

8:30	8:45	Introduction Overview of IG Programs	Vickie Johncox
8:45	9:00	BIONET Resource	Dave Kristofferson
9:00	9:45	Operating System, File Structure	Vickie Johncox
9:45	10:30	Electronic Mail	Kathy Berg
10:30	10:45	Break	
10:45	12:00	Sequence Entry	Vickie Johncox
12:00	1:00	Lunch	
1:00	2:00	Sequencing Gel Management	Spencer Yeh
2:00	3:15	NA and Peptide Sequence Analysis	
3:15	3:30	Break	
3:30	5:00	Problems	

Friday, July 20

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Vickie Johncox
10:30	10:45	Break	
10:45	12:00	Collecting Sequences, Pattern Searching	
12:00	1:00	Lunch	
1:00	3:15	Sequence Similarity Searches	Sunil Maulik
3:15	3:30	Break	
3:30	4:30	Review and Questions	Vickie Johncox

**An Introduction to the IntelliGenetics Suite
Course Schedule
September 15 and 16 1988**

Thursday, September 15

8:30 8:45	Introduction	Vickie Johncox
8:45 9:00	Introduction to BIONET	Dave Kristofferson
9:00 9:30	Operating System, File Structure	Trish Benton
9:30 10:00	Electronic Mail	Kathy Berg
10:00 10:15	Organization of IntelliGenetics	Murray Summers
10:15 10:30	Break	
10:30 11:15	Sequence Data Entry and Editing	Karen Davis
11:15 12:00	Sequencing Gel Management	Trish Benton
12:00 1:00	Lunch	
1:00 3:15	NA and Peptide Sequence Analysis	Karen Davis
3:15 3:30	Break	
3:30 5:00	Problems	

Friday, September 16

8:00 9:00	Open Classroom	Consultants
9:00 10:30	Database Structure and Sequence Retrieval	Trish Benton
10:30 10:45	Break	
10:45 12:00	Collecting Sequences, Pattern Searching	
12:00 1:00	Lunch	
1:00 3:15	Sequence Similarity Searches	Sunil Maulik
3:15 3:30	Break	
3:30 4:00	Problems	
4:00 4:15	Review and Questions	Vickie Johncox

**An Introduction to the IntelliGenetics Suite and BIONET
Course Schedule
November 17 and 18, 1988**

Thursday, November 17

8:30	9:00	Introduction	Vickie Johncox
9:00	9:15	Introduction to BIONET	Dave Kristofferson
9:15	9:45	Operating System, File Structure	Vickie Johncox
9:45	10:15	Electronic Mail	
10:15	10:30	BREAK	
10:30	11:15	Sequence Data Entry and Editing	
11:15	12:00	Sequencing Gel Management ,	
12:00	1:00	LUNCH	
1:00	1:30	Problems	
1:30	3:30	NA and Peptide Sequence Analysis	Karen Davis
3:30	3:45	BREAK	
3:45	5:00	Problems	
5:00	6:30	BREAK	
6:30	9:00	BIONET-specific topics	BIONET staff

**An Introduction to the IntelliGenetics Suite and BIONET
Course Schedule
November 17 and 18, 1988**

Friday, November 18

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Karen Davis
10:30	10:45	BREAK	
10:45	12:00	Collecting Sequences, Pattern Searching	Karen Davis
12:00	1:00	LUNCH	
1:00	3:00	Sequence Similarity Searches	Vickie Johncox
3:00	3:15	BREAK	
3:15	3:45	Problems	
3:45	4:00	Review and Questions	Vickie Johncox
4:00	5:00	BIONET-specific topics	BIONET staff

**Introduction to BIONET
Thursday Evening Schedule
November 17, 1988**

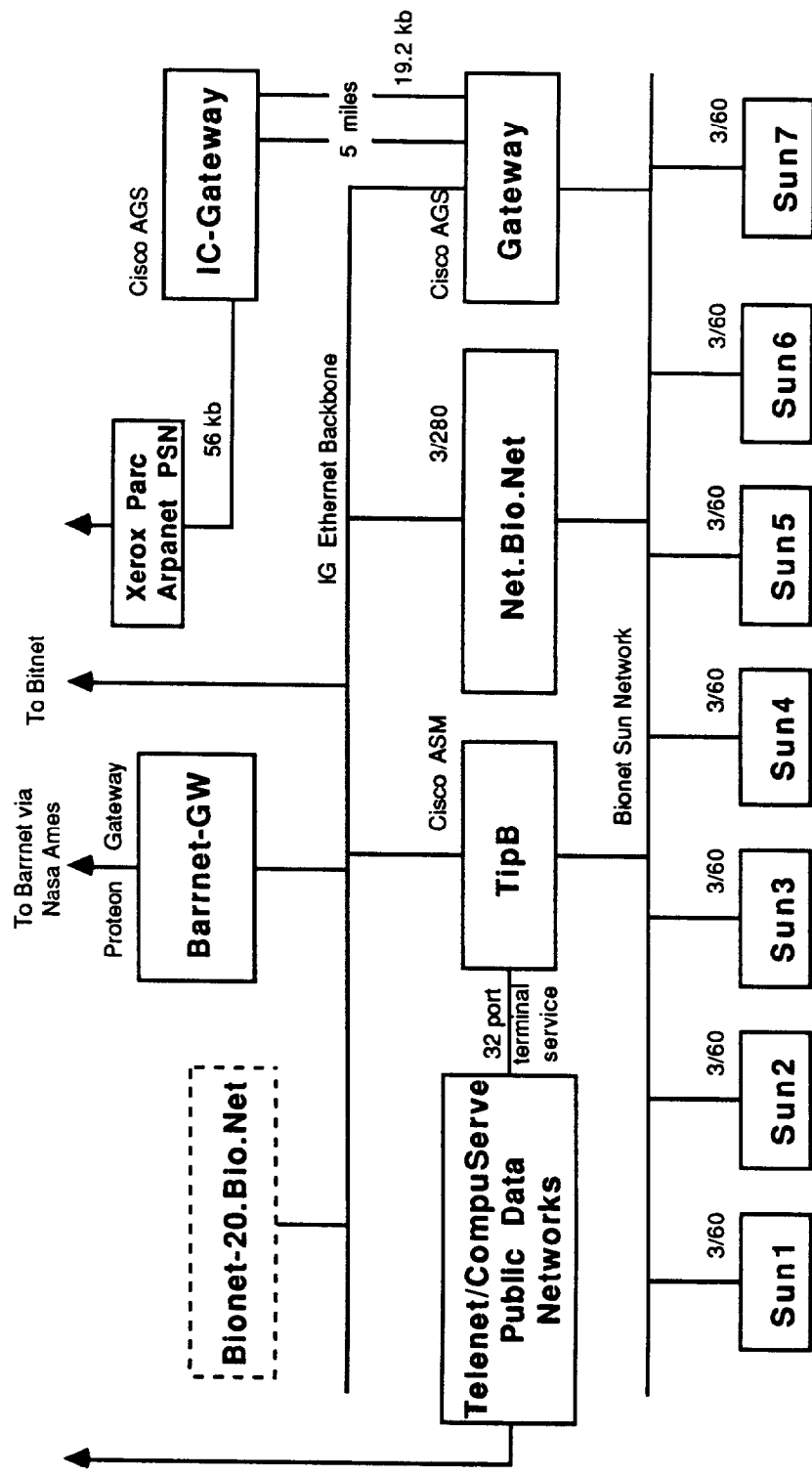
6:30	7:15	Connecting to BIONET, File Transfers	David Kristofferson
7:15	7:45	TOPS-20 Operating System, EMACS Text Editor	Karen Davis
7:45	8:00	BREAK	
8:00	8:30	Bulletin Boards, E-mail Addresses, Help Me	Karen Davis
8:30	9:00	Other BIONET Software	Spencer Yeh

**Introduction to BIONET
Friday Afternoon Schedule
November 18, 1988**

4:00	5:00	FASTA-MAIL	Spencer Yeh
------	------	------------	-------------

VII. BIONET Computer Facilities

A diagram of the BIONET computer facilities follows on the next page.



Bionet Central Computing Resource

VIII. Testimonials

Copies of two recently received testimonial letters are included in this section.

HARVARD MEDICAL SCHOOL
DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY

Tel. (617) 732- 2046
Fax: 738-0516
Internet: Hirsh@BIONET-20.bio.net



25 Shattuck Street
Boston, Massachusetts 02115

Jay Hirsh
Assoc. Professor
260 Longwood Ave.

October 21, 1988

Re: Statement in support of NIH grant to the BIONET computer facility.

To whom it may concern:

We have been members of the BIONET computer facility for approximately 3 years. We have found this facility to be of great importance to our research. We make extensive use of the homology searching and sequence analysis routines, and are just beginning to take full advantage of the database searching routines. These analyses have uncovered a number of promoter elements of the *Drosophila* dopa decarboxylase gene (*Ddc*) that are conserved through evolution and are functionally important ((Scholnick et al (1986) *Science* 234, 998-1002; Bray & Hirsh (1986) *EMBO J.*, 5,2305-2312; Bray et al (1988) *EMBO J.*,7,177-188.). We enclose a copy of this last reference. Even though it did not utilize directly the Bionet resource, this manuscript shows directly that one of the initially identified conserved elements is a CNS-specific regulatory element.

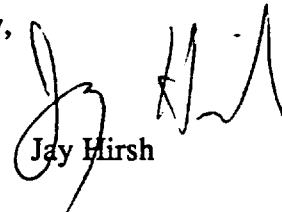
We have recently extended this analysis to a cell-specific enhancer of *Ddc* (Johnson, McCormick, Bray & Hirsh, in preparation), an ~800 bp segment that contains a number of functional elements. This approach is also proving to be valid in this region, in that a number of conserved elements are showing interesting *in vivo* functions in regulating different aspects of the neuronal pattern of expression of *Ddc*.

Within the past year, we have also begun to make extensive use of the Bionet E-mail facility. We now routinely communicate with European and American colleagues using this system. Given the cost of international phone calls, our use of this facility alone would pretty much justify the yearly bionet fee.

We continue to use this facility, even though there is presently a local computer facility claiming to offer comparable services at a comparable cost. This decision is due to the superb level of support that we have derived from the Bionet staff: Questions regarding operation of the system are answered expertly and promptly. The programs and program manuals have been continually evolving, such that the programs are now accessible to even novice users in the lab. When I have made queries-such as wondering whether database searches couldn't be done without hanging on the phone for hours- I have been surprised and amazed that the staff have gone to what appear to be major efforts to implement such searching programs. Our experiences with the aforementioned local computer facility, where the level of staff support was almost nill, have shown the value of such services.

In summary, I hope that NIH will continue to support the Bionet resource. This program has attracted a dedicated and skilled staff that provides services and support that we have not been able to find elsewhere. An interruption in the funding of this program would certainly cause disruption to a large number of users who are highly dependent on these people and this system.

Sincerely,



Jay Hirsh

THE JOHNS HOPKINS UNIVERSITY

BALTIMORE, MARYLAND 21218

DEPARTMENT OF BIOLOGY

November 28, 1988

Dr. David Kristofferson
BIONET
c/o Intelligenetics
700 E. El Camino Road
Mountain View, CA 94040

Dear Dave,

I am addressing this letter to you personally because I dislike the "To Whom It May Concern" format - it makes me feel as though I am writing to an answering machine. Please feel free however to show the contents to anyone.

There are several statements that I would like to make about BIONET, its services, its staff and the Intelligenetics suite of DNA and protein sequence analysis software that they offer.

1. BIONET offers a level of support for the researcher that cannot be duplicated by many universities or research centers.

While DNA/protein sequence analysis software can be put on the university or medical school mainframe computer or run on personal computers, the problem has always been using it: training people in fundamentals (try to find understandable documentation on some of these programs) and then teaching them the significance of the variables involved and the limitations of the algorithms used. BIONET has made a considerable effort to address these problems. They offer affordable training at Mountain View, California and are willing to come to you if they have staff and time. They have made a real attempt to write complete and understandable and useful documentation for the system and the programs. Most important is their online help and on-call systems experts who answer immediate questions about programs and data. It is difficult for a systems operator at a university computer center to be well-versed on all the software running on the machine. Even if they are, they need the time to explain the programs. These overworked systems operators must rely on local experts or bright students to help people with questions about individual programs. Molecular biologists need easily accessible, experienced professionals to answer their questions. BIONET offers those professionals.

2. The BIONET staff are highly competent and extremely cooperative. I have nothing but praise for the personnel at BIONET and the service and support they offer. Over the last year, I have received advice (good advice, by the way), had problems solved, and talked about on-site training. Everyone who dealt with me was knowledgeable, professional and very helpful. I really appreciate having these people on call to answer my questions, which they do with remarkable rapidity and skill.

3. The systems analysts and programmers at BIONET listen to the users. Several times I have had questions about the running of a program or suggestions about format or documentation and the people at BIONET always listened and corrected the program's problem or explained to me my problem. They are always improving the system and it is wonderful never to be stuck with a "quirk" in the program.

4. Access to BIONET is very easy for the computer novice. As the computer "expert" in Dr. Roseman's laboratory, I find that people are very cautious about accessing the mainframe computer on campus. Mistakes cost money and since the system is not dedicated to one suite of software, the interface has to be more complicated. Most people don't want to learn DOS for the PC let alone a whole new operating system for the Vax or the IBM. BIONET allows the novice easy access while retaining all the system commands for the more expert or more daring among us. People in my laboratory are connecting to the BIONET system with ease and learning as they work. This is an immense time-saver for me since the online help keeps them from calling my name every five minutes and distracting me from my own research.

5. BIONET provides communication lines between scientists that are easily used and therefore frequently used. For a while BIONET was the only large scale bulletin board system for molecular biologists. Now more and more communications lines are open, but the problem is again getting people to use them. Some of them require Bitnet or Arpanet, etc. and that requires people to access their mainframe computers and learn how to run electronic mail systems that are not as user-friendly as the BIONET system. BIONET is now acting as a gateway to a number of these systems and I for one, am grateful that I can hear the news, or send the news without running one more program.

I am very pleased with BIONET and the Intelligenetics programs. They have enabled my colleagues and me to further our work in molecular biology by providing the sequence analysis software in a form that we could learn and understand. I believe that BIONET provides necessary and unique services for the research scientist.

Very truly yours,



Donna K. Fox, Ph.D.
Associate Research Scientist